AFRL-OSR-VA-TR-2013-0100

Novel Spectro-Temporal Codes and Computations for Auditory Signal Representation and Separation

Kumaresan

University of Rhode Island

FEBRUARY 2013
Final Report

AIR FORCE RESEARCH LABORATORY
AF OFFICE OF SCIENTIFIC RESEARCH (AFOSR)/RSL
ARLINGTON, VIRGINIA 22203
AIR FORCE MATERIEL COMMAND

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
10-01-2013	Final	03/01/2009 to 08/31/2012

4. TITLE AND SUBTITLE

Novel Spectro-Temporal Codes and Computations for Auditory Signal Representation and Separation

5a. CONTRACT NUMBER

5b. GRANT NUMBER

FA9550-09-1-0119

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Ramdas Kumaresan

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Rhode Island
Kelley Hall, 4 East Alumni Avenue
Kingston RI 02881

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFOSR
875 N Randolph St
Arlington, VA 22203
Dr. Willard Larkin/RSL

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

AFRL-OSR-VA-TR-2013-0100

12. DISTRIBUTION/AVAILABILITY STATEMENT

Distribution A: Approved for public release

13. SUPPLEMENTARY NOTES

14. ABSTRACT

In the past three years, we have developed algorithms that emulate the phenomenon of "synchrony capture" in the auditory nerve. Synchrony capture means that the dominant component in a given frequency band preferentially drives auditory nerve fibers innervating the entire corresponding frequency region of the cochlea. Our algorithm, called the synchrony capture filterbank (SCFB) consists of a bank of broadly tuned filters (not unlike the basilar membrane) in cascade with narrower filters (not unlike outer hair cells) that adaptively lock onto locally-dominant frequency components to produce synchrony capture. This local behavior enables a robust encoding of the running power spectrum based on relative numbers of channels recruited by different frequency components. The filterbank precisely tracks individual time-varying frequency components, such as low harmonics and formant frequencies in speech, in the midst of noise and auditory clutter. This precise tracking in turn can be used to enhance the separation of concurrent periodic sounds. . We envision that the project will result in improved front-ends that can enhance voices in noise and better separate

15. SUBJECT TERMS

Auditory-Inspired Signal Processing, Synchrony Capture, Speech processing by the auditory system.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Ramdas Kumaresan
U	U	U	U		19b. TELEPHONE NUMBER *(Include area code)* 617-406-9882

Reset

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Synchrony capture filterbank (SCFB): Auditory-inspired signal processing for tracking individual frequency components in speech

Ramdas Kumaresan[a)] and Vijay Kumar Peddinti

Department of Electrical,

Computer and Biomedical Engineering,

University of Rhode Island,

Kingston,

RI 02881

Peter Cariani

Department of Otology & Laryngology,

Harvard Medical School,

Boston,

MA 02114

(Dated: December 28, 2012)

Abstract

A processing scheme for speech signals is proposed that emulates synchrony capture in the auditory nerve. The role of stimulus-locked spike timing is important for representation of stimulus periodicity, low frequency spectrum, and spatial location. In synchrony capture dominant single frequency components in each frequency region impress their time structures on temporal firing patterns of auditory nerve fibers (ANFs) with nearby characteristic frequencies (CFs). At low frequencies, for voiced sounds, synchrony capture divides the nerve into discrete CF territories associated with individual harmonics. An adaptive, synchrony capture filterbank (SCFB) consisting of a fixed array of traditional, passive linear (gammatone) filters cascaded with a bank of adaptively tunable, bandpass filter triplets is proposed. Differences in triplet output envelopes steer triplet center frequencies via voltage controlled oscillators (VCOs). The SCFB exhibits some cochlea-like responses, such as two-tone suppression and distortion products, and possesses many desirable properties for processing speech, music, and natural sounds. Strong signal components dominate relatively greater numbers of filter channels, thereby yielding robust encodings of relative component intensities. The VCOs precisely lock onto harmonics most important for formant tracking, pitch perception, and sound separation.

PACS numbers: 43.72 Ar, 43.64 Bt, 43.64 Sj

Synchrony capture filterbank (SCFB): Auditory-inspired signal processing for tracking individual components in speech

I. INTRODUCTION

For the past three decades there has been significant interest in developing computational signal processing models based on the physiology of the cochlea and auditory nerve (AN)[1]. The hope has been that artificial systems can be designed and built using signal processing strategies gleaned from nature that can equal or exceed human auditory performance. Our work in this area is motivated by neurophysiological observations of the synchrony capture phenomenon in the auditory nerve that were originally reported by Sachs et al.[2] and Delgutte et al.[3]. This paper proposes such a biologically-inspired signal processing strategy for processing speech and audio signals.

If one systematically examines the temporal representation of low harmonics of complex sounds in the auditory nerve, synchrony capture is a striking feature. Synchrony capture means that the dominant component in a given frequency band preferentially drives auditory nerve fibers innervating the entire corresponding frequency region of the cochlea[3]. Here, virtually all fibers innervating this cochlear place region, i.e. those with CFs in the vicinity of the frequency of the dominant component, synchronize exclusively to the dominant component, in spite of the presence of other nearby weaker components that may be closer to their CFs. At moderate and high sound pressure levels, fibers spanning an entire octave or more of CF are typically driven at their maximal rates and exhibit firing patterns related to a single, dominant component in each formant region. Because of the symmetric nature of cochlear tuning, this dominant component mostly drives fibers whose CFs lie above it in frequency. Figures 1 and 2 provide examples of this phenomenon in slightly different forms. Figure 1a shows peristimulus time histograms (PSTHs) for a five-formant synthetic vowel

[a)]Department of Electrical, Computer and Biomedical Engineering, University of Rhode Island, Kingston RI 02881; Electronic address: kumar@ele.uri.edu

sound. Sharp boundaries characteristic of synchrony capture are seen between the different CF regions driven by different dominant, formant-region harmonics of the multi-formant vowel. Note that in Figure 1a other non-dominant harmonics in the vowel formant regions are not explicitly represented.

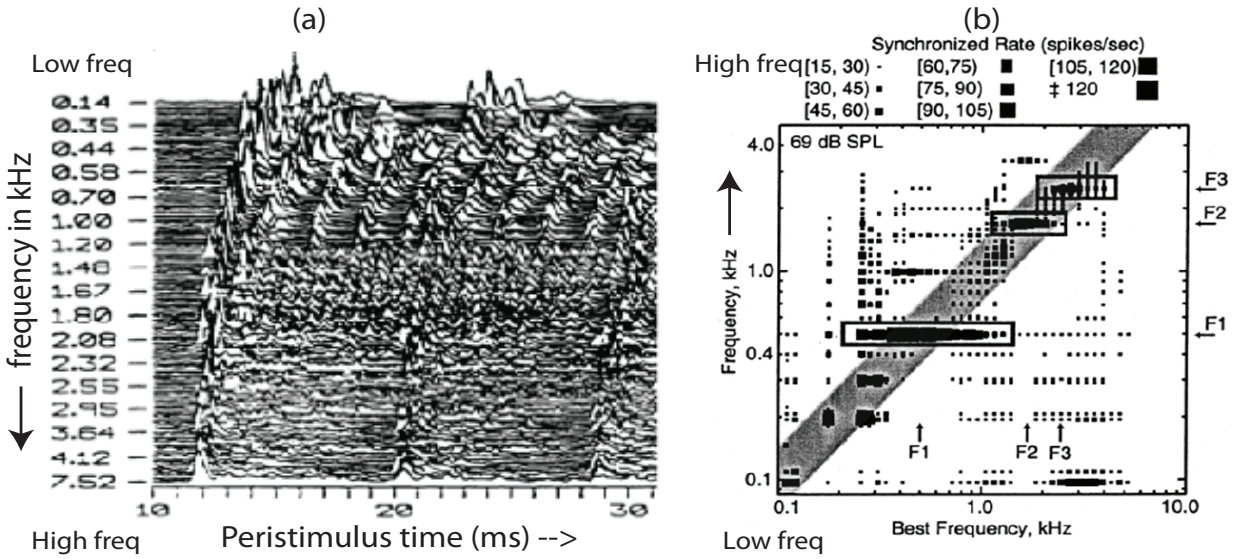

FIG. 1. Two views of the representation of vowel-like sounds in the AN. a) Peristimulus time histograms for cat ANF arranged by characteristic frequency in response to the onset of a five-formant synthetic vowel (/da/) reprinted from Seeker-Walker and Searle (1990)[4]. (b) Distribution of synchronized rates in ANFs in response to a standard vowel /da/ with three formants F_1, F_2, and F_3. F_0 =100Hz. Reprinted from Sachs et al. (2002)[5].

Figure 1b summarizes temporal firing patterns observed in the cat auditory nerve in response to a three-formant synthetic vowel[5]. Relative synchronized rates of fibers to different component frequencies are shown as a function of fiber characteristic (CF) or best frequency (BF). Sizes of squares indicate synchronized rates (larger squares = higher rates). The diagonal gray band shows regions where temporal firing periodicities match fiber BFs, and the dark horizontal swaths indicate capture of fibers over a range of fiber best frequencies by individual stimulus components. The most prominent swaths are the synchrony capture regions for the dominant harmonics associated with each of the three formants (en-

closed boxes). In addition to capture by dominant harmonics in formant regions, low-CF fibers show synchrony to less-intense, non-formant, low harmonics (n=1-3) when frequencies of those harmonics happen to be near their respective CFs (dark boxes within the gray diagonal band).

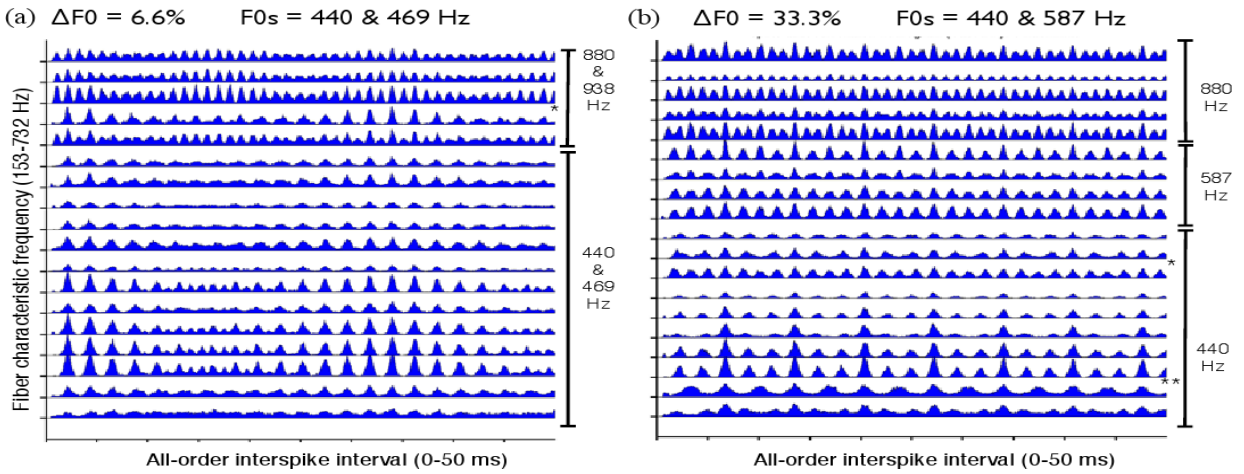

FIG. 2. Synchrony capture of adjacent partials for two frequency separations. The two neurograms show all-order interspike interval distributions for individual cat auditory nerve fibers as a function of CF in response to complex tone dyads presented 100 times at 60 dB SPL. Each tone of the pair consisted of equal amplitude harmonics 1-6. New analysis of dataset originally reported in Tramo et al. (2001)[6]. (a) Responses to a tone dyad a musical minor second apart (16:15, ΔF_0=6.6%). Vertial bars indicate CF regions where one predominant interspike interval pattern predominates. The CFs of the fibers shown are: 153, 283, 309, 345, 350, 355, 369, 402, 402, 431, 451, 530, 588, 602, 631, 660, 724, and 732 Hz. Misordered interval patterns (single-asterisked histograms) are likely due to small CF measurement errors. (b) Response to a tone dyad a musical fourth apart (4:3, ΔF_0=33.3%). Three distinct interspike interval patterns associated with individual partials (440, 587, and 880 Hz) are produced in different CF bands, with abrupt transitions between response modes. One fiber shows locking to distortion product $2f_1 - f_2$ near its CF (double-asterisked histogram, $2f_1 - f_2$ = 293 Hz, CF = 283 Hz). Fiber CFs were 153, 283, 346, 350, 355, 369, 402, 402, 431, 451, 530, 588, 602, 631, 660, 662, 724, 732, and 732 Hz.

Synchrony capture is most directly apparent when distributions of all-order interspike intervals (spike autocorrelation histograms) produced by individual fibers are plotted as a function of fiber CF (cochlear place)[7]. Figure 2 shows fiber interspike interval patterns in response to two concurrent complex harmonic tones (n= 1-6). For a stimulus in which pairs of harmonics are close together (Figure 2a, ΔF_0= 6.6% of F_0), all of the fibers in the region synchronize to the composite, modulated waveform. In this case, the temporal firing patterns in the whole CF region follow the beating of the adjacent partials, producing low-frequency fluctuations in firing rate that are associated with perceived roughness[6]. Here, when the adjacent partials are sufficiently close together there are no separate temporal, interspike interval representations of individual harmonics themselves. On the other hand, for a tone pair for which the lower harmonics are relatively well separated in frequency (Figure 2b, $\Delta F_0 = 33.3\%$ of F_0), different CF regions are captured by one or another partial. Thus each harmonic component drives a discrete region of the cochlea in which its temporal pattern dominates, with almost no zones of beating (right panel, there are different CF zones with different interval peak patterns). The result is that each individual partial has its own swath of auditory nerve fibers that produce corresponding interspike interval patterns.

The foregoing examples indicate that auditory nerve fibers synchronize preferentially to dominant components in the signal. In signal processing terms the peripheral auditory system appears to treat these dominant components as "carrier" frequencies. The effects of the weaker surrounding components (other harmonics) then manifest themselves as modulations on these carriers (as can be seen in Figure 1a).

A. Significance of synchrony capture

Synchrony capture may have implications for neural representations of periodicity and spectrum, as well as for F_0-based sound separation and grouping. Synchrony capture in the auditory nerve permits representation of relative intensity that is level-invariant, and thus is useful for representing the normalized power spectrum in a robust manner. The num-

bers of fibers locking onto particular frequency components give indications of the relative intensities of the corresponding components. This is a robust means of encoding their relative magnitudes using neural elements with limited dynamic ranges. The proposed SCFB algorithm[8] attempts to emulate this behavior using adaptive filters to create a competition for channels amongst frequency components that not only accurately reflects their relative magnitudes, but is also invariant with respect to absolute signal amplitude.

This signal processing strategy for encoding relative intensities has relevance for auditory nerve representations. Global temporal representations of lower-frequency sounds in the auditory nerve, called population-interval distributions or summary autocorrelations, implicitly utilize such principles to represent pitch and timbre (e.g. vowel formant structure)[7,9–11]. The most direct signal processing analogues of these global temporal auditory nerve models are the ensemble interval histograms (EIHs)[12]. Essentially, dominant frequency components below 5 kHz that are present at any given instant partition the cochlear CF territory into swaths of auditory nerve fibers (ANFs) that have similar temporal discharge patterns (and hence similar interval distributions). In the context of global population-interval representations that sum together interspike intervals across the entire auditory nerve, relative intensities of partials are conveyed through relative numbers of all-order interspike intervals associated with their respective locally-dominant components rather than numbers of CF channels recruited. Whether through relative numbers of pooled intervals or of similarly-responding channels, this parcellation of the cochlea into competing synchronization zones efficiently utilizes the entire auditory nerve for signal representation.

Synchrony capture could also potentially be utilized by place-based brainstem auditory representations that analyze excitation boundaries by using local across-CF comparisons of temporal firing patterns[13]. Here the abrupt temporal pattern discontinuities associated with synchrony capture increase contrast and the precision of boundary estimations in such coding schemes.

Further, synchrony capture may facilitate F_0-pitch formation and sound separation by enhancing temporal representations of individual, resolved harmonics at the expense of those

produced by interactions of multiple, unresolved harmonics. Synchrony capture has the effect of minimizing periodicities related to beatings of adjacent harmonics, as can be seen in the lack of composite interspike interval patterns when the harmonics are well separated (Figure 2b). The temporal auditory nerve representation of a harmonic complex with low, well-separated harmonics thus resembles a series of interspike interval patterns each of which resembles that of a pure tone of corresponding frequency.

The enhancement of the representation of individual harmonics in turn has implications for F_0-based sound separation. Most acoustic signals in everyday life are mixtures of sounds from multiple sources. In order to separate multiple concurrent sounds, human listeners mainly rely on differences in onset times and fundamental frequencies F_0s. Results of psychophysical experiments suggest that separation of multiple auditory objects with different fundamentals, such as those produced by multiple voices or musical instruments, crucially depends on the presence of perceptually-resolved harmonics (n<5)[14]. These resolved harmonics dominate in pitch perception and have high pitch salience[15].

In terms of interspike interval representations of individual partials (as seen in Figure 2), the effect of synchrony capture is to separate the interspike interval patterns of adjacent partials if they are separated by more than some threshold ratio, or to fuse them together if they are not. It is therefore not unreasonable to hypothesize that the synchrony capture process might play a role in whether adjacent partials are fused together or separated perceptually. For frequencies for which there is significant phase-locking, synchrony capture behavior thus qualitatively parallels tonal separations and fusions that are associated with harmonic resolution and critical bands. These parallels notwithstanding, the size of psychophysically-measured critical bandwidths in cats, roughly twice those of humans, cast some doubt on a simple, direct correspondence[16].

The mechanism in the auditory pathway whereby the harmonically-related components of each of two concurrent harmonic complexes fuse together to produce two F_0-pitches at their respective fundamentals is not yet understood. The two F_0-pitches can be heard out, even if the harmonics of the two complexes are interleaved, provided that the unrelated, ad-

jacent harmonics are sufficiently separated in frequency. In this context, synchrony capture minimizes temporal patterns associated with interactions between adjacent, harmonically-unrelated partials, thus eliminating interaction products that might otherwise degrade the representations of the individual harmonics and hinder their grouping and separation on the basis of shared interspike intervals.

For the above reasons, it seems reasonable to emulate synchrony capture in a signal processing algorithm.

B. Design rationale for the SCFB algorithm

Although the explicit goal of the SCFB is to emulate synchrony capture in the auditory nerve and not to model cochlear biophysics, because its signal processing design was partially inspired by cochlear structure, some discussion of the latter is useful in understanding the former. A schematic of the proposed SCFB algorithm is shown in Figure 3a. It consists of a bank of K fixed, relatively broad filters in cascade with tunable, narrower filters that produce the synchrony capture behavior. This nesting of broad and narrow filters is not unlike coarse and fine gradations in a vernier scale. Tuning of the adaptive filters is carried out via frequency discriminator loops (FDLs) on time scales of milliseconds to tens of milliseconds, making real-time frequency tracking possible.

In any attempt to reverse-engineer biological auditory functions, it is useful to consider artificial systems that exhibit behaviors not unlike their natural counterparts. The phenomenon of synchrony capture appears similar to the well known "frequency capture" behavior of traditional FM receivers such as FM discriminators and phase lock loops. Frequency capture[17] occurs when an FM receiver locks on to a strong FM signal even in the presence of other interfering, relatively weaker FM signals. One such FM receiver circuit is a frequency discriminator[18](p.206), which uses stagger-tuned bandpass filters whose output envelopes are differenced to obtain the demodulated baseband signal. Such circuits are known to exhibit frequency capture. The signal processing architecture proposed here was

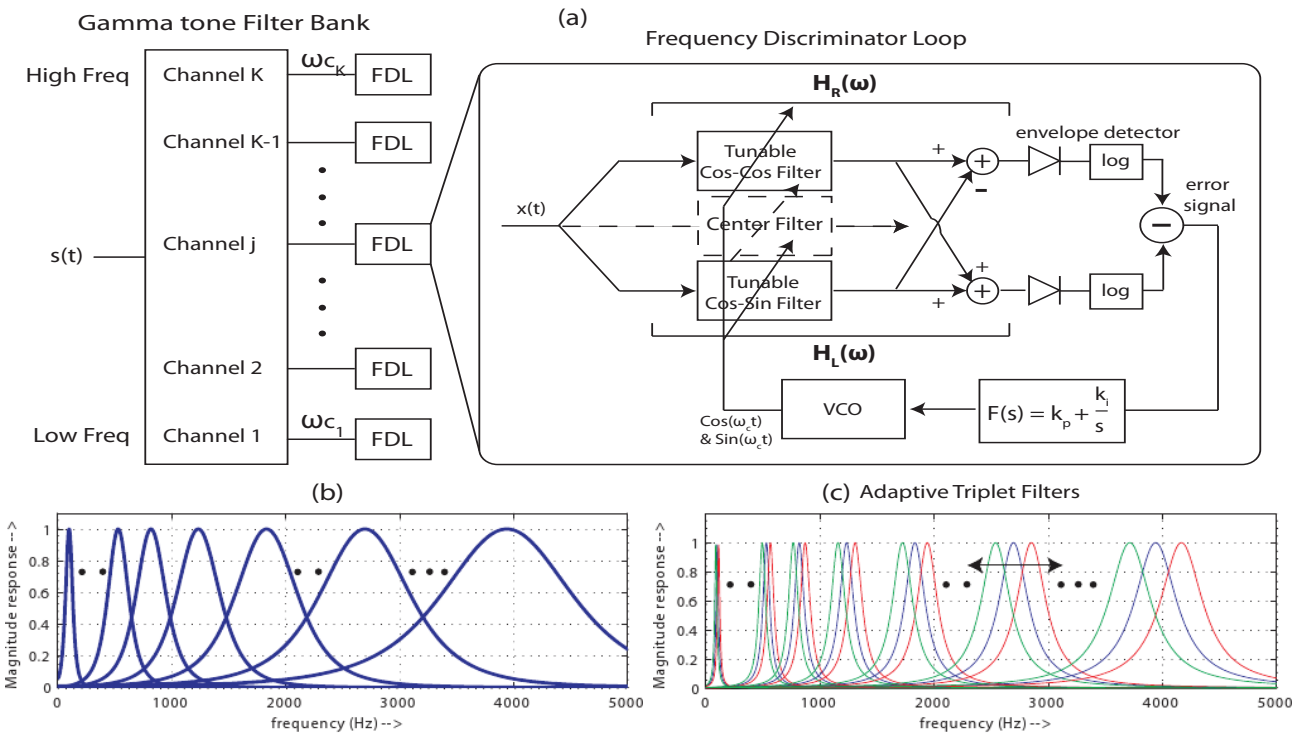

FIG. 3. Synchrony capture filterbank (SCFB). (a) The filterbank architecture consists of K constant-Q gammatone filters whose logarithmically-spaced center frequencies span the desired audible frequency range. Each filterbank channel consists of a frequency discriminator loop (FDL) cascaded with each of the K gammatone filters. The output of each channel, $y_c(t)$, is obtained from its center filter. See sections II and III for details. Frequency responses of fixed and tunable filters in the SCFB. Bottom left panel (b) shows the frequency responses of fixed gammatone filters (the black dots indicate that not all filter responses are shown). Bottom right panel (c) shows the Frequency responses of the tunable bandpass filter (BPF) triplets that adapt to the incoming signal. One BPF triplet is associated with each fixed filter, such that coarse filtering of the fixed gammatone filters is followed by additional, finer filtering by tunable filters. The nested arrays of fixed, coarse and adjustable, fine filters are arranged in a manner similar to a vernier scale.

designed with both these circuits and possible cochlear analogues in mind.

In the SCFB architecture, the fixed gammatone filterbank with relatively coarse bandpass tunings (Q = 4) emulates the behavior of the passive basilar membrane whose stiffness

decreases monotonically from base to apex. The bandwidths of the gammatone filters were chosen to approximate cochlear impulse responses and tuning characteristics observed for input signals at high sound pressure levels and are thought to be consequences of largely passive mechanical filtering[19]. In the SCFB architecture, finer frequency tuning is achieved using a second layer of narrower bandpass filters (BPFs, Q=8) that emulate the filtering functions of outer hair cells (OHCs). In the cochlea, while inner hair cells (IHCs) are thought to be relatively passive mechanoelectrical transducers, outer hair cells also have active electromechanical processes that permit them to change length under the influence of their transduction currents, thereby amplifying local mechanical vibrations[20].

The proposed adaptive bandpass filter (BPF) triplets that form the heart of the frequency discriminator loop (FDL) consist of three relatively narrowly tuned filters with slightly offset center frequencies that are in cascade with each fixed filter of the passive gammatone filterbank. This arrangement contrasts with the situation in the cochlea, where OHCs with their active processes and narrower tunings are in bidirectional interaction with the more broadly tuned motions of the basilar membrane[19]. The BPF triplets are locally adaptive and are tuned based on differences in amplitudes of signals output by the filters in the triplet. Although broadly similar designs were available in the adaptive filtering literature[21,22], independent of auditory modeling, it was the spatial arrangement of outer hair cells (OHCs) observed in mammalian cochleae[23] that inspired this particular triplet design. The lateral amplitude differencing process in each BPF triplet amounts to taking the spatial derivative of the local amplitude spectrum at that particular cochlear location. Such lateral differencing processes could conceivably be carried out over time spans of up to tens of milliseconds via lateral interactions in intracochlear and olivocochlear neural networks[24] (p.15, Fig.1.13 (A)),[25],[26](p.289, Fig.11).

The tuned, oscillatory motility of outer hair cells inspired use of a voltage-controlled oscillator (VCO) to tune the filter triplets. Feedback control of triplet tuning could also be potentially implemented via other signal processing mechanisms. The action of hair cell stereocilia that open ion channels preferentially in one direction suggests half-wave rectifi-

cation of the signal, an operation similar to envelope detection that is already commonly used in auditory modeling. The nonlinear response characteristics of hair cells inspired the logarithmic compression of the envelope (see section II.B) that is used by the frequency discriminator loop to capture dominant signals and suppress weaker ones. All of these design features stem from the general idea that many aspects of cochlear function and auditory nerve behavior can be emulated by frequency tracking circuits.

C. Organization of the paper

This paper first describes the operation of components of the adaptive filters, followed by the architecture of the SCFB as a whole. In section II, FDLs and their use as basic tone followers are presented. As mentioned earlier, each FDL is made up of three tunable bandpass filters (called a BPF triplet). Tuning of the triplet filters is effected using voltage controlled oscillators (VCOs). In section II.A a simple tone follower (STF) consisting of a BPF triplet and a VCO is described that is capable of tracking the frequency of a tone. The linear equivalent circuit of the tone follower is presented, which is useful in choosing the loop filter parameters of the FDL. The dominant tone follower (DTF) is then developed in section II.B. The DTF uses a simple nonlinearity in the feedback loop of the FDL to lock on to the dominant tone when the input consists of more than one tone. In other words, the DTF is capable of synchrony capture. In section II.C a practical implementation of the BPF triplet is presented that has several desirable characteristics for signal processing purposes, such as linear phase, perfect even and odd symmetry and a single VCO operation.

In section III a traditional fixed gammatone filterbank is combined in cascade with a bank of FDLs to form the synchrony capture filterbank (SCFB). Responses of the filterbank to harmonic tone complexes, isolated vowels, and running speech are presented in section IV. Correspondences with cochlear response characteristics and auditory nerve behavior are discussed in section V. The section V also includes relationships of the proposed algorithm to previous research.

II. TONE FOLLOWERS AND FREQUENCY CAPTURE

Frequency discriminator loops (FDLs) have been used for synchronizing transmitter and receiver oscillators in digital and analog communication systems for decades[27,28]. Typically, in a communication receiver, an FDL brings the receiver oscillator frequency close to the transmitter frequency, i.e., within the lock-in range of a phase lock loop, such that it can lock the two oscillators[29]. The structure of the frequency tracking algorithms used here, called tone followers, are similar to the FDLs used in communication systems. The block diagram of a generic FDL is shown in Figure 4. It consists of a frequency error detector (FED), a loop filter and a voltage controlled oscillator (VCO). The FED outputs an error signal $e(t)$ that is proportional to the difference between the frequency of the input signal ω_1 and the frequency of the VCO output, ω_c. The loop filter provides the control voltage to the VCO and drives its frequency such that $\omega_c - \omega_1$ tends to zero. Typically, the system function $F(s)$ of the the loop filter determines its dynamics and has the form $k_p + k_i/s$ where k_p and k_i are the proportional and integral gain factors[30], respectively (more details below in Section II.A).

Section II.A, describes how an FDL is used as a simple tone follower (STF) and defines its components. A linear equivalent circuit of the FDL is also provided. In most realistic sound processing contexts one encounters multiple sinusoidal signals (as in a voiced speech formant). In section II.B, a dominant tone follower (DTF) is described that is capable of following a dominant tone in the presence of other interfering weaker tones and exhibits synchrony capture. This is realized by using a compressive nonlinearity in the feedback path. The linear equivalent circuit for DTF is essentially identical to that of the STF.

A. A simple tone follower (STF)[22]

The frequency discriminator loop (FDL) (Figure 4) tracks the frequency of an input tone by using a frequency error detector (FED) that steers the center frequencies of the VCOs of the triplet adaptive filters (Figure 5). Another type of FED is described in Appendix A.

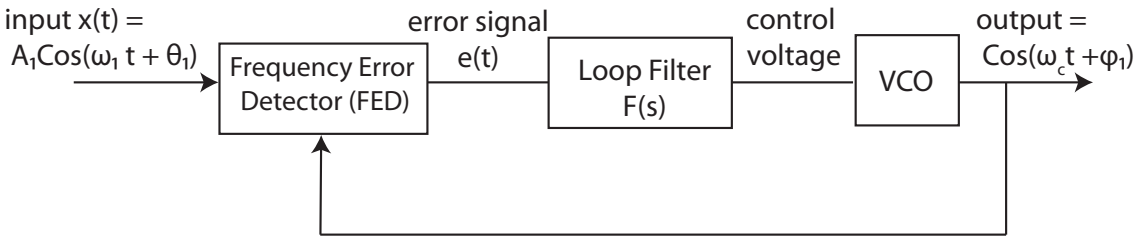

FIG. 4. A generic frequency discriminator loop (FDL). The error signal e(t) is a measure of the frequency difference between the input signal and the VCO. See Figures 5 and 8 for details of specific frequency error detectors.

In principle, the FED consists of three identically shaped tunable band pass filters (BPFs), $H_R(\omega)$, $H_C(\omega)$ and $H_L(\omega)$, initially centered around frequencies $\omega_c + \Delta$, ω_c and $\omega_c - \Delta$, respectively. The subscripts R, C and L stand for the right, center and left filters, respectively. As ω_c, the frequency of the VCO (in Figure 4) is changed, the center frequencies of the BPFs' also change accordingly, such that these filters' response functions slide along the frequency axis. The spacing between triplet filters (Δ) is fixed. Only the left and right filters are used in calculating the error signal $e(t)$. The envelope detectors compute the (squared) envelope of the BPFs' outputs. When a tone, $A_1 \cos(\omega_1 t + \theta_1)$ is presented to the FED, the average values of the (squared) envelopes for right and the left filters are $e_R(t) = |A_1 H_R(\omega_1)|^2$ and $e_L(t) = |A_1 H_L(\omega_1)|^2$, respectively. (If the input tone's frequency changes with time then e_R and e_L are also functions of time t.) Then the error signal $e(t)$ is computed as the ratio of the difference of the envelopes $(e_R(t) - e_L(t))$ to their sum $(e_R(t) + e_L(t))$.

Note that the ratio eliminates the amplitude of the input signal A_1 from $e(t)$, and now $e(t)$ is just related to the frequency error $\omega_c - \omega_1$. Instead of computing the ratio, an AGC circuit at the input could have been used to normalize the amplitude. The principle is to move the frequency responses of the BPFs $H_R(\omega)$ and $H_L(\omega)$ (and $H_C(\omega)$) in tandem, under the control of the VCO frequency ω_c, such that when the error $e(t) = 0$, ω_c equals ω_1. So, the VCO tracks the input frequency.

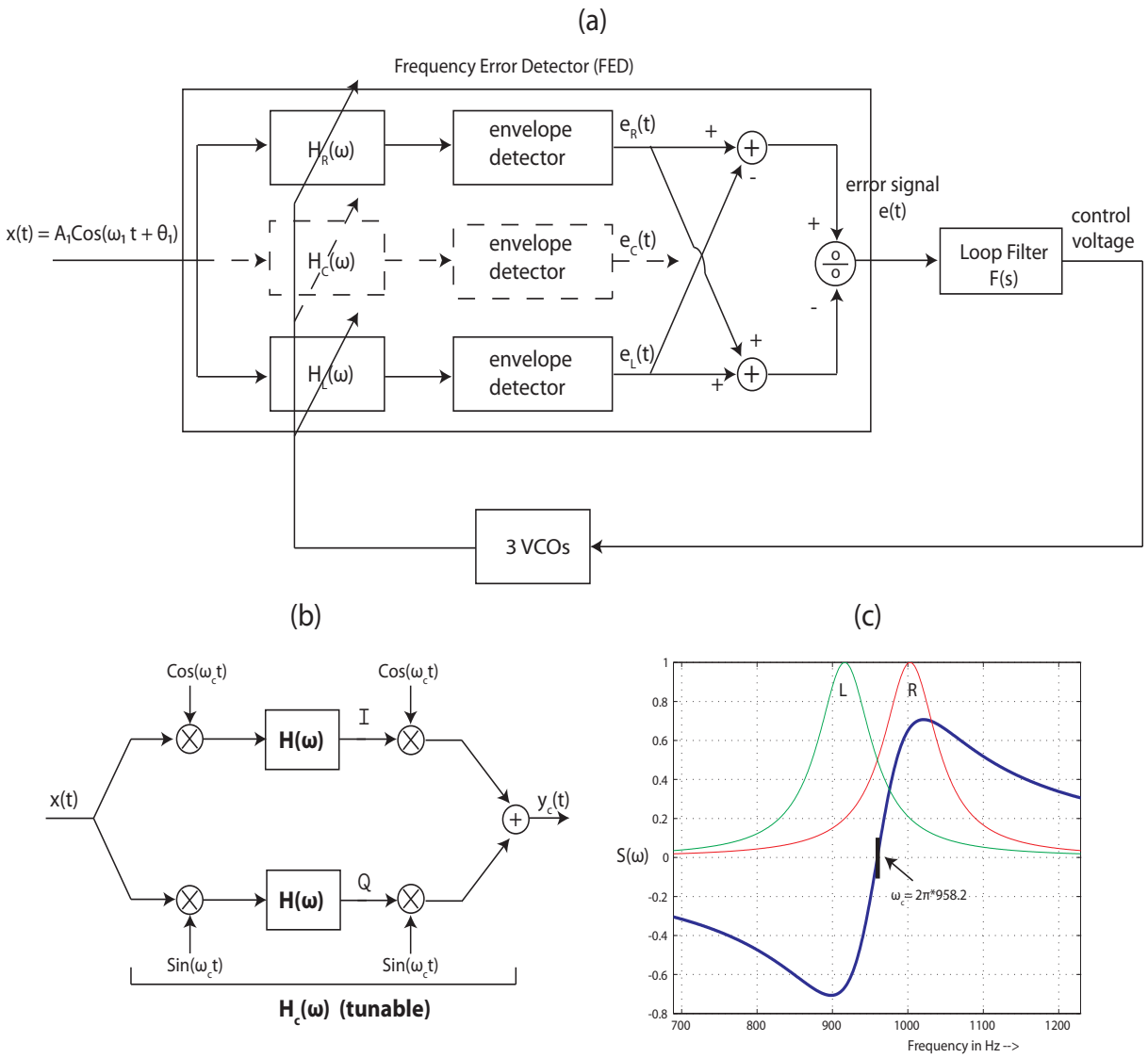

FIG. 5. Frequency error detector (FED) used in the simple tone follower (STF). Error signal $e(t)$ is computed using the formula $\frac{e_R(t)-e_L(t)}{e_R(t)+e_L(t)}$. The envelopes $e_L(t)$, $e_R(t)$, and $e_C(t)$, are obtained as I^2+Q^2. The I and Q for center filter $H_C(\omega)$, are the outputs of the LPFs shown in (b). $H_L(\omega)$ and $H_R(\omega)$ have the same structure but with oscillator frequencies at $\omega_c-\Delta$ and $\omega_c+\Delta$ respectively. The discriminator transfer characteristics $S(\omega)$ (thick line) and magnitude responses of left and right filters (thin lines) are shown in (c).

The frequency discriminator function $S(\omega)=\dfrac{|H_R(\omega)|^2-|H_L(\omega)|^2}{|H_R(\omega)|^2+|H_L(\omega)|^2}$ (also called the "S-curve"[29]), is shown in Figure 5c. When a tone $A_1\cos(\omega_1 t+\theta_1)$ is applied as the input, then

$e(t) = S(\omega_1)$. In the interval $\omega_c - \Delta < \omega < \omega_c + \Delta$ the error voltage $e(t)$ is approximately linear, so $e(t) \approx k_s(\omega_c - \omega_1)$. k_s is called the frequency discriminator constant[29].

The tunable BPFs are built using the filter structure shown in Figure 5b (called "cos-cos" structure), which shows how $H_C(\omega)$ (centered at ω_c) is realized using two lowpass filters (LPFs). Identical LPFs with frequency response $H(\omega)$ are sandwiched between two multipliers in both the lower and upper branches of the circuit. Both the multipliers in the upper branch are supplied with $\cos \omega_c t$ (hence the name cos-cos structure) and the lower branch are supplied with a $\sin \omega_c t$ from the same VCO with frequency ω_c. It can be easily shown that,

$$H_C(\omega) = H(\omega + \omega_c) + H(\omega - \omega_c). \tag{1}$$

Similarly, the BPF $H_L(\omega)$ (or $H_R(\omega)$) is implemented as a cos-cos structure with the same LPF filters but with the VCO frequency at $\omega_c - \Delta$ (or $\omega_c + \Delta$). Together the three filters shown inside the FED box in Figure 5a is called a *BPF triplet*. The frequency spacing between these filters, Δ, is kept fixed. Only the left and right filters are used in calculating the error signal $e(t)$.

The center filter envelope is used to declare a "track" condition, i.e. that the filter has converged on a tonal input. When this convergence occurs at the input tone frequency ω_1, then the envelope of the center filter output $e_C(t)$ will satisfy the following condition,

$$e_L(t) = e_R(t) = \mu e_C(t) \tag{2}$$

for some constant μ. If the filter shapes are chosen such that $|H_R(\omega_c)| = |H_L(\omega_c)| = 0.707|H_C(\omega_c)|$ (i.e., 3-dB points of the right and left filter coincide with the center frequency of the center filter), then $\mu = 0.5$. If the above condition is satisfied, then the input is a tone whose frequency coincides with the VCO frequency ω_c, and a "track" condition is declared. Such channel outputs can be used to compute the pitch frequency of a complex tone. This FED structure requires three VCOs operating at $\omega_c - \Delta$, ω_c and $\omega_c + \Delta$ to realize the $H_L(\omega)$, $H_C(\omega)$, and $H_R(\omega)$ respectively.

An approximate linear equivalent circuit of the frequency discriminator loop can provide

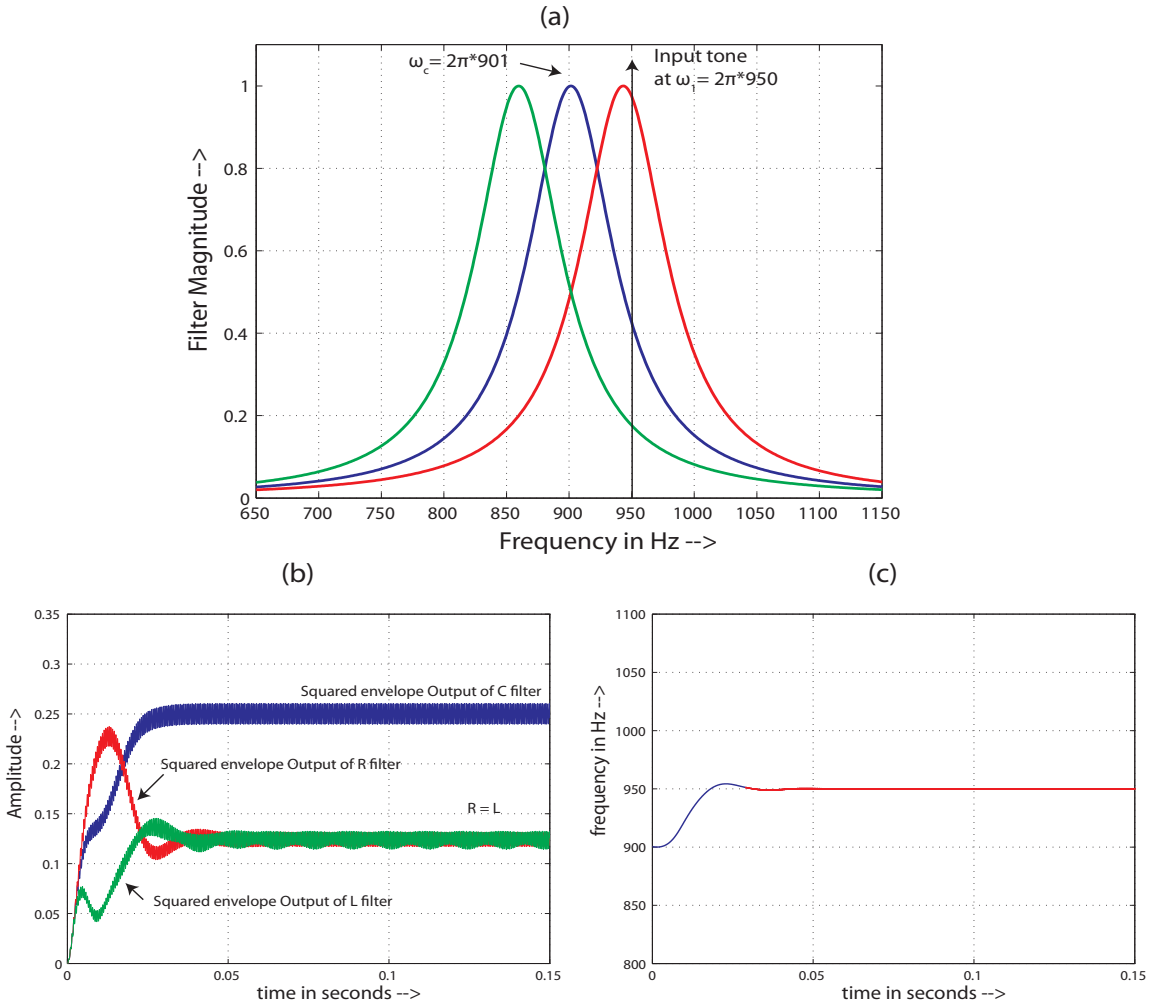

FIG. 6. Convergence of a BPF triplet on an input tone at ω_1. (a) Frequency responses of BPF triplet filters in relation to an input tone. The input tone frequency is $\omega_1 = 2\pi\times950$ Hz. Initially the L, C, and R filters are centered at $\omega_c - \Delta = 2\pi\times859$ Hz, $\omega_c = 2\pi\times901$ Hz and $\omega_c + \Delta = 2\pi\times943$ Hz, respectively. Since initially $\omega_1 > \omega_c$, the initial envelope output $e_R(t)$ is greater than $e_L(t)$, so the normalized error $e(t)$ is positive. This positive value of $e(t)$ causes the VCO frequency ω_c to increase until ω_c equals ω_1. (b) Time course of envelopes e_L(t), e_C(t) and e_R(t). Note that the envelopes $e_R(t)$ and $e_L(t)$ become equal after some settling time and that $e_C(t)$ reaches a higher plateau, where e_L(t)=e_R(t)=0.5e_C(t). (c) VCO frequency track for the C filter.

some insight into the behavior of the tone follower (Figure 7). Here the input tone and the oscillator output are replaced by their frequency values ω_1 and ω_c, respectively. Recall that the frequency error detector (FED) outputs a voltage level proportional to the frequency difference $\omega_1 - \omega_c$. Therefore, the FED in Figure 5a is modeled by a proportionality constant k_s. Assuming that we operate the discriminator loop in the region $\omega_c - \Delta < \omega < \omega_c + \Delta$, this constant k_s is the gain factor representing the slope of the S-curve shown in Figure 5c. Assuming that the sandwiched LPF in Figure 5b has a system function $1/(s+\alpha)$, where α represents its 3-dB bandwidth, it can be shown that the frequency error discriminator constant k_s is equal to $2\Delta/(\Delta^2+\alpha^2)$ (see Appendix B). In addition, note that the calculation of the envelopes needed to estimate the frequency difference entails a group delay τ_g. This time delay is represented by its Laplace transform $e^{-s\tau_g}$ in Figure 7. At low frequencies the BPF filters are narrower, and hence τ_g is relatively large. At high frequencies $\tau_g \approx 0$. In Figure 7, $e^{-s\tau_g}$ is approximated (using Padé approximation[31]) by a ratio of first order s-polynomials,

$$e^{-s\tau_g} \approx \frac{1-\gamma s}{1+\gamma s} \tag{3}$$

where $\gamma = \tau_g/2$. The controller is a loop filter whose transfer function is $F(s) = k_p + k_i/s$ where k_p is the proportional constant and k_i is the integral constant ([30], page 254).

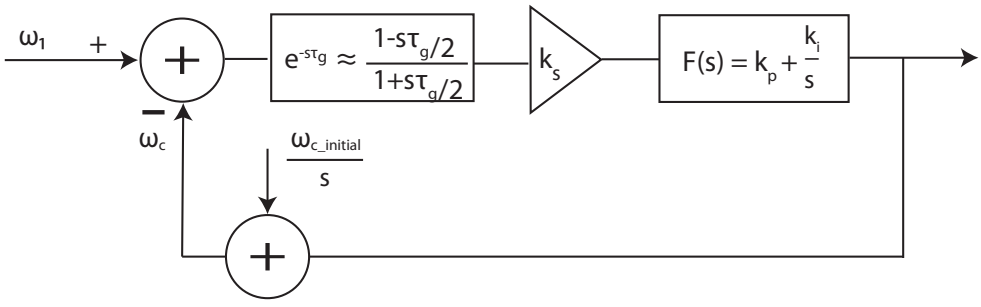

FIG. 7. Linearized model of the frequency discriminator loop.

Then, the closed loop transfer function $H(s)$ of the linearized model is

$$H(s) = B(s)/A(s) \tag{4}$$

$$= \frac{\frac{1-\gamma s}{1+\gamma s} k_s \left(k_p + \frac{k_i}{s}\right)}{1 + \frac{1-\gamma s}{1+\gamma s} k_s \left(k_p + \frac{k_i}{s}\right)} \tag{5}$$

After some simplification we find that the denominator polynomial $A(s)$, which determines the settling time τ_s of the loop, is given by the following expression,

$$A(s) = s^2 + \frac{(1 + k_s k_p - \gamma k_s k_i)}{(\gamma - \gamma k_s k_p)} s + \frac{k_i k_s}{(\gamma - \gamma k_s k_p)} \tag{6}$$

Using Routh's Stability Criterion, the conditions for stability are given by

$$(\gamma - \gamma k_s k_p) > 0 \Rightarrow k_p < \frac{1}{k_s}$$

$$(1 + k_s k_p - \gamma k_s k_i) > 0 \Rightarrow \gamma k_i - k_p < \frac{1}{k_s}$$

$$k_i k_s > 0 \Rightarrow k_i > 0, (k_s \text{ is positive})$$

We need to find k_p and k_i such that the step response has a desirable settling time. This is done using the standard pole positioning method ([30], page 233) based on Bessel polynomials. For a second order system with a normalized settling time of 1 second, the Bessel roots of the closed loop system are at $-4.05 \pm j2.34$. And for a desired settling time of τ_s seconds, the roots are scaled by τ_s, i.e., $(-4.05 \pm j2.34)/\tau_s$. Hence the corresponding Bessel polynomial is $s^2 + (8.11/\tau_s)s + 21.90/\tau_s^2$. By comparing this polynomial with the $A(s)$ in Eq. 6 , we can write the following two linear equations in terms of k_p and k_i:

$$a_1 k_i + b_1 k_p = c_1$$

$$a_2 k_i + b_2 k_p = c_2$$

where

$$a_1 = \tau_s \gamma k_s \qquad b_1 = -k_s (\tau_s + 8.11\gamma) \qquad c_1 = (\tau_s - 8.11\gamma)$$

$$a_2 = \tau_s^2 k_s \qquad b_2 = 21.90\gamma k_s \qquad c_2 = 21.90\gamma$$

Solving for k_p and k_i obtains

$$k_p = \frac{1}{k_s}\frac{\beta - 1}{\beta + 1},$$
$$k_i = \frac{1}{k_s}\left(21.90\frac{\gamma}{\tau_s^2}\right)\frac{2}{\beta + 1}, \tag{7}$$

where $\beta = 8.11\left(\frac{\gamma}{\tau_s}\right) + 21.90\left(\frac{\gamma}{\tau_s}\right)^2$.

An example of the operation and convergence dynamics of a simple tone follower (STF) in response to a pure tone nearby in frequency is illustrated in Figure 6, and described in the caption. The step response of the linear equivalent circuit (step size is $950 - 901 = 49$ Hz) coincides almost exactly with that of the frequency track shown in Figure 6c.

B. Dominant tone follower (DTF)

The simple tone follower (STF) is suitable for tracking one tone, but in real world acoustic environments, pure tonal signals are only rarely encountered. Instead, the vast majority of signals are mixtures of complex sounds from multiple sources that can contain nearby partials or harmonics. Here a dominant tone follower (DTF) is needed that can track the frequency of a dominant partial in a signal even in the presence of other interfering ones, similar to the synchrony capture behavior observed in the auditory nerve. A simple modification of the STF described above that employs a nonlinearity in the feedback loop results in the dominant tone follower (DTF) described below.

Consider a signal $x(t)$ consisting of a tone at frequency $\omega_1 = 2\pi f_1$ and an interfering tone at $\omega_2 = 2\pi f_2$.

$$x(t) = A_1\cos(\omega_1 t + \theta_1) + A_2\cos(\omega_2 t + \theta_2) \tag{8}$$

Let us assume that $A_1 > A_2$, i.e., the tone at ω_1 is dominant. We rewrite $x(t)$ using complex notation as follows.

$$x(t) = \Re\{A_1 e^{j(\omega_1 t + \theta_1)}(1 + \frac{A_2}{A_1}e^{j\Delta\omega t + j\Delta\theta})\} \tag{9}$$

where $\Re$ stands for "Real part of", $\Delta\omega = \omega_2 - \omega_1$ and $\Delta\theta = \theta_2 - \theta_1$, and $j = \sqrt{-1}$. Since $A_2/A_1 < 1$, (using the approximation that $e^y \approx 1 + y$ for $y < 1$, in the above expression) we

have,

$$x(t) \approx a(t)\cos(\phi(t)), \tag{10}$$

where the envelope is

$$a(t) \approx e^{\log A_1 + \frac{A_2}{A_1}\cos(\Delta\omega t + \Delta\theta)}, \tag{11}$$

and the phase function is

$$\phi(t) \approx \omega_1 t + \theta_1 + \frac{A_2}{A_1}\sin(\Delta\omega t + \Delta\theta). \tag{12}$$

The derivative of $\phi(t)$ (i.e., the instantaneous frequency (IF)[18], p. 180) and the log-envelope are as follows:

$$\frac{d\phi(t)}{dt} \approx \omega_1 + \frac{A_2}{A_1}\Delta\omega\cos(\Delta\omega t + \Delta\theta), \tag{13}$$

$$\log a(t) \approx \log A_1 + \frac{A_2}{A_1}\cos(\Delta\omega t + \Delta\theta). \tag{14}$$

The symbol log denotes natural logarithm. Note that the average value of IF is ω_1, the dominant tone's frequency, and similarly, the average value of the log-envelope is the dominant tone's log amplitude. Either of these properties can be utilized for frequency discrimination purposes. An exact expression for the log-envelope of $x(t)$ can also be obtained as follows:

$$a^2(t) = |A_1 e^{j\omega_1 t + j\theta_1} + A_2 e^{j\omega_2 t + j\theta_2}|^2 = A_1^2 + A_2^2 + 2A_1A_2\cos(\Delta\omega t + \Delta\theta). \tag{15}$$

Taking logarithm and using the infinite series expansion for $\log(1+x)$ we have

$$\log a(t) = \log A_1 + \sum_{n=1}^{\infty} \frac{1}{n}\Big(\frac{A_2}{A_1}\Big)^n \cos(n\Delta\omega t + n\Delta\theta). \tag{16}$$

Note that Eq. 14 retains only the first term in the infinte sum above. Also note that the average value of $\log a(t)$ is $\log A_1$. On the other hand, the average value of the squared envelope $a^2(t)$ is $(A_1^2 + A_2^2)$.

A frequency discriminator can lock on to ω_1 by filtering the instantaneous frequency (IF, assuming that it is available) using a low-pass filter (LPF) with a cut off frequency $\Delta\omega$. Alternatively, the log-envelope can also be used to capture the dominant signal (Figure 8). In an FDL the logarithmically compressed envelope signal, $\log a(t)$, can be low pass filtered

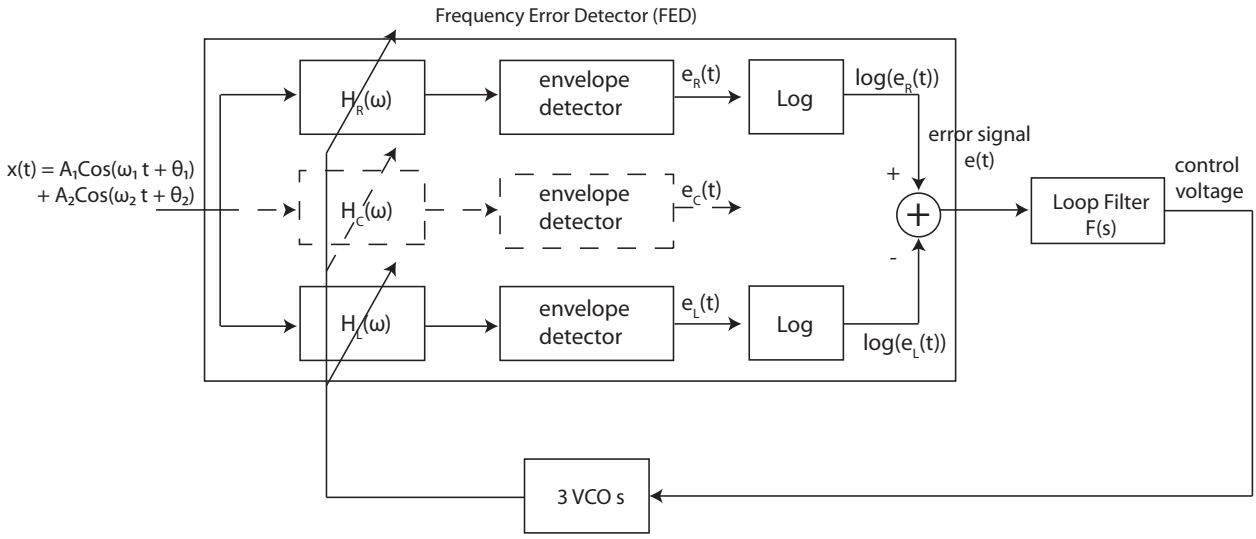

FIG. 8. Frequency error detector (FED) for the dominant tone follower (DTF). The error signal $e(t)$ is computed using the formula $\log\left(\frac{e_R(t)}{e_L(t)}\right)$.

(with the same cut off frequency, $\Delta\omega$, as in the case of IF) to obtain $\log A_1$. This can then be used to lock on to the dominant tone in the input.

Compared to the simple tone follower, note that the envelopes in the dominant tone follower are now compressed using a logarithmic nonlinearity before they are low pass filtered (by the loop filter). If the input is just one tone ($x(t) = A_1\cos(\omega_1 t + \theta_1)$) then the corresponding smoothed squared envelopes at the outputs of the right ($H_R(\omega)$) and left ($H_L(\omega)$) filters are $A_{1R}^2 = A_1^2|H_R(\omega_1)|^2$ and $A_{1L}^2 = A_1^2|H_L(\omega_1)|^2$ respectively. So, the error signal is $e(t) = 2\log(A_{1R}/A_{1L})$. Note that $e(t)$ is proportional to the frequency difference $\omega_1 - \omega_c$ and does not depend on the amplitude A_1 (as in STF).

Now, consider the case of an input $x(t)$ with two tones as in Eq. 8. Then, there are two cases. In the first case, assume that the same tone (either at ω_1 or ω_2) dominates both (right and left) filters' outputs. Then, clearly the (average) error is $2\log(A_{1R}/A_{1L})$ or $2\log(A_{2R}/A_{2L})$ depending on which tone dominates. Since the loop tends to drive this error to zero, the VCO frequency ω_c changes such that the left and right filter's log-amplitudes are equal. Thus ω_c tends to track the dominant tone. In contrast, if the nonlinearity is absent

then the left and the right filters produce (squared, averaged) envelopes equal to $A_{1L}^2 + A_{2L}^2$ and $A_{1R}^2 + A_{2R}^2$, which result in ω_c settling in between ω_1 and ω_2, i.e., no capture. Thus, the compressive non-linearity helps steer the VCO to the dominant signal's frequency.

In the second case, if the tone at ω_1 dominates the left filter output and the tone at ω_2 dominates the right filter output, then the error $e(t)$ is proportional to $\log(A_{2R}/A_{1L})$ and the VCO frequency is adjusted by the loop such that $A_{2R} = A_{1L}$. That is ω_c averages in between ω_1 and ω_2. In summary, if one tone is sufficiently bigger than the other, then capture occurs, but if two tones are close in frequency and have equal or almost equal amplitudes, then the VCO locks on to a weighted average frequency. This behavior is similar to that seen in the auditory nerve (Figure 2b) for nearby partials.

The linear equivalent circuit for the DTF is essentially identical to that of the STF developed in section II.A, except that the parameter k_s is slightly different $\left(k_s = \dfrac{4\Delta}{\Delta^2 + \alpha^2}\right)$ (see Appendix B). Figure 9 shows an example of a DTF homing in on a stronger tone in the presence of a nearby weaker tone (vertical arrows). Such dominant tone followers are used as the building blocks for the proposed filterbank algorithm described below in section III.

C. A practical implementation of the frequency discriminator loop (FDL)

This section presents the design of an FDL which incorporates a single VCO and matched BPF triplet filters. This implementation of the BPF triplet (and the FDL) that requires only one VCO has several advantages over those described above. The filters that form the BPF triplet are implemented as linear phase filters. The BPF triplet is implemented with the help of odd/even prototype filters such that they result in perfectly matched, symmetrical, left ($H_L(\omega)$) and right ($H_R(\omega)$) filters. That is, their frequency response magnitudes are exactly equal at the VCO's frequency ω_c. Further, the computation of the envelopes $e_R(t)$ and $e_L(t)$ does not explicitly require in-phase (I) and quadrature phase (Q) signal components. Instead the envelope is simply obtained by taking the absolute value of the signal, i.e. , the full-wave-rectified output, and low-pass filtering it. The three bandpass filters that constitute

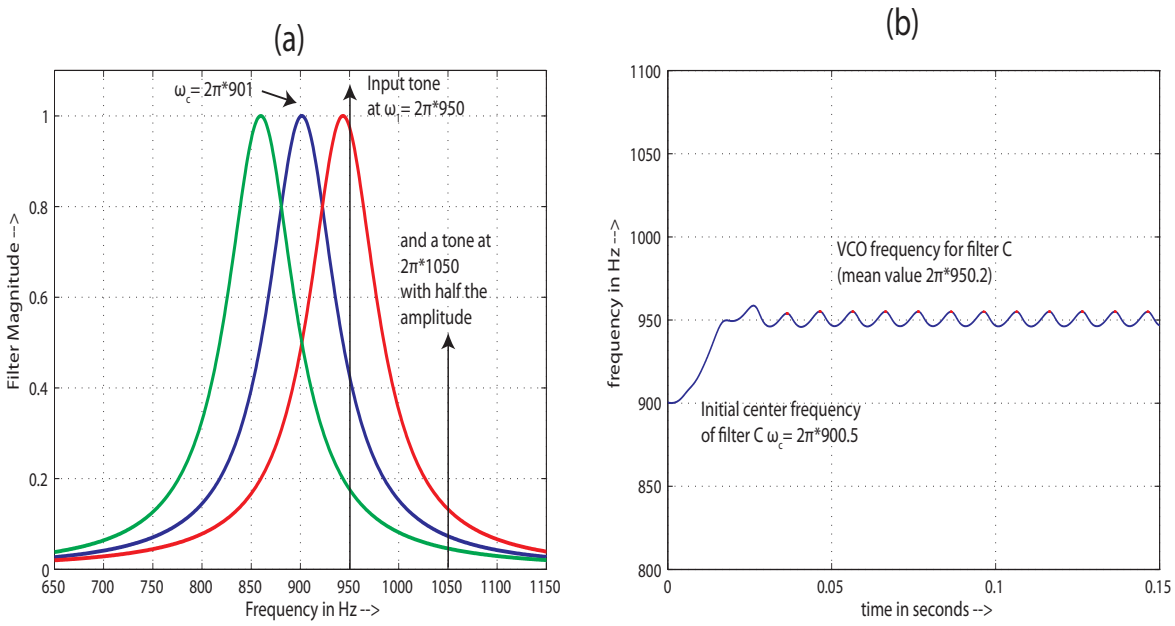

FIG. 9. Behavior of a DTF in response to two nearby tones of different amplitude. (a) Frequency response of BPF triplet filters and the input tones (vertical arrows, dominant tone at $\omega_1 = 2\pi\times950$ Hz, plus a half-amplitude interfering tone at $\omega_1 = 2\pi\times1050$ Hz. (b) Track of the VCO frequency for the center filter C. With minor fluctuations, the VCO tracks the stronger 950 Hz tone in-spite of the weaker 1050 Hz interferer.

the BPF triplet can all be synthesized from a single prototype noncausal, low-pass impulse response,

$$h(t) = e^{-\alpha|t|}, \tag{17}$$

$$H(\omega) = 2\alpha/(\omega^2 + \alpha^2). \tag{18}$$

Any other even impulse response function with unimodal low pass frequency response characteristics (such as, $h(t) = e^{-\beta t^2}$) can also be used as a prototype filter. Let $h_1(t)$ and $h_2(t)$ represent the impulse responses of frequency translated filters, given by

$$h_1(t) = e^{-\alpha|t|}\cos\Delta t, \text{ and } h_2(t) = e^{-\alpha|t|}\sin\Delta t, \tag{19}$$

where Δ is the translation frequency. So,

$$H_1(\omega) = (H(\omega - \Delta) + H(\omega + \Delta))/2,$$
$$H_2(\omega) = j(H(\omega - \Delta) - H(\omega + \Delta))/2, \tag{20}$$

where $j = \sqrt{-1}$. Δ is chosen equal to α, so that Δ is the 3-dB point of $H(\omega)$. The frequency responses $H_1(\omega)$ and $H_2(\omega)$ are purely real and imaginary, respectively.

$H_1(\omega)$ and $H_2(\omega)$ are embedded as part of the tunable band pass filters $G_1(\omega)$ and $G_2(\omega)$ shown in Figures 10a and 10b, respectively. $G_1(\omega)$ is called a cos-cos filter (same structure as Figure 5b) and $G_2(\omega)$ is named a cos-sin filter.

$$G_1(\omega) = (H_1(\omega - \omega_c) + H_1(\omega + \omega_c))/2,$$
$$G_2(\omega) = j(H_2(\omega - \omega_c) - H_2(\omega + \omega_c))/2. \tag{21}$$

The frequency responses $G_1(\omega)$ and $G_2(\omega)$ are both real and even and are shown in Figure 10c. These frequency responses can be tuned by changing ω_c.

Assume for the moment, that the systems $H_1(\omega)$ and $H_2(\omega)$ sandwiched between the multipliers are identical. Then, note that the system functions of a generic cos-cos structure, $G_1(\omega)$, and cos-sin structure, $G_2(\omega)$, are related by the expression $G_2(\omega) = jsgn(\omega)G_1(\omega)$ for sufficiently large ω_c. That is, cos-sin structure has an additional term which signifies a Hilbert transform when compared to cos-cos structure. This stems from the fact that the multipliers in the upper/lower branches of Figure 10b are cosine and sine unlike the cos-cos filter in Figure 10a. This is a seemingly new way of realizing a band-pass Hilbert transformer. The outputs of the cos-cos and cos-sin filters are then added/subtracted (see Figure 11) to obtain the overall right/left filter responses $H_R(\omega)$ and $H_L(\omega)$ (Figure 10d), respectively. That is,

$$H_R(\omega) = G_1(\omega) - G_2(\omega), \text{ and } H_L(\omega) = G_1(\omega) + G_2(\omega). \tag{22}$$

Substituting for $G_1(\omega)$ and $G_2(\omega)$ in Eq. 22 from Eq. 21, we have,

$$H_R(\omega) = (H_1(\omega - \omega_c) + H_1(\omega + \omega_c))/2 + j(H_2(\omega - \omega_c) - H_2(\omega - \omega_c))/2,$$
$$H_L(\omega) = (H_1(\omega - \omega_c) + H_1(\omega + \omega_c))/2 - j(H_2(\omega - \omega_c) - H_2(\omega - \omega_c))/2. \tag{23}$$

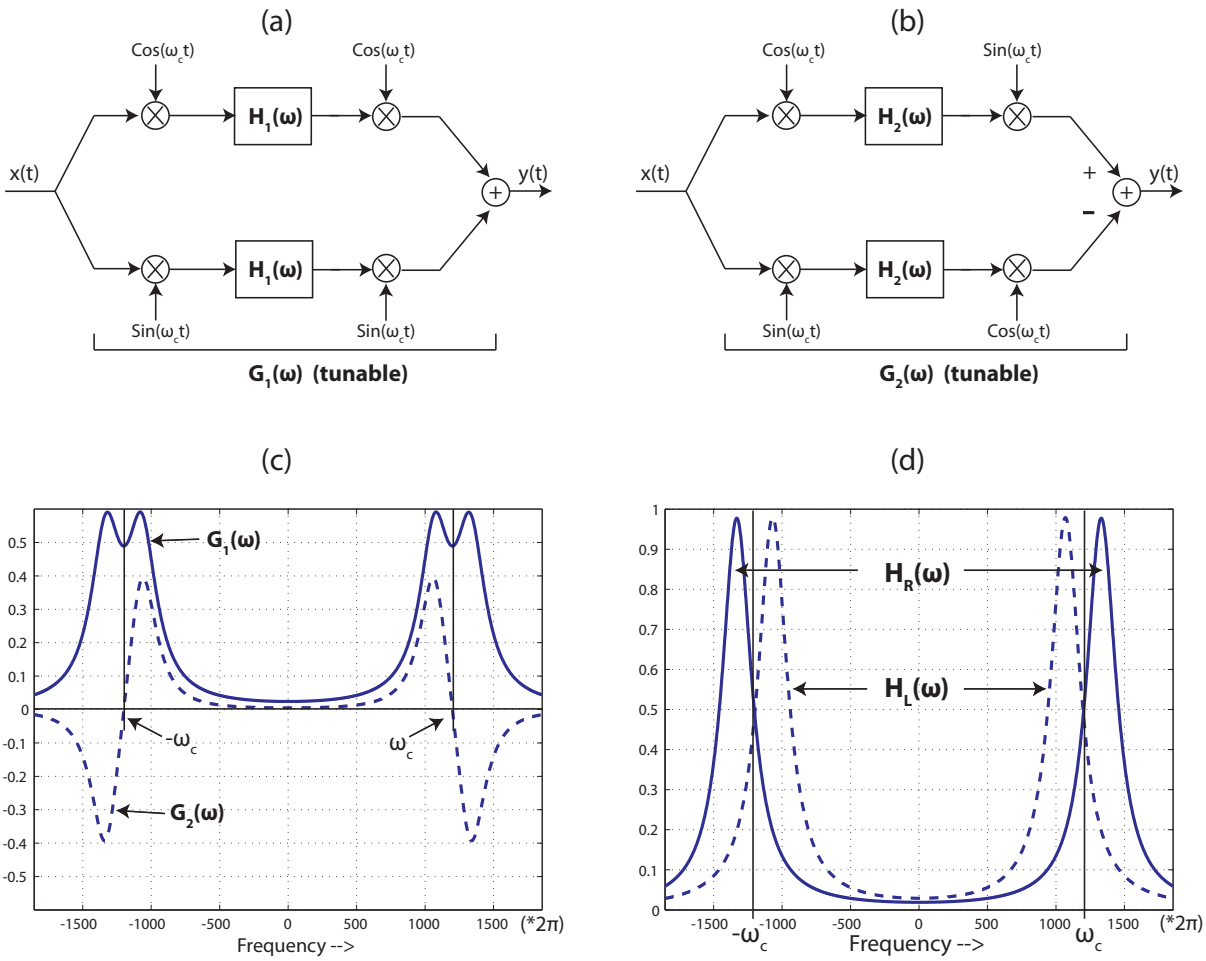

FIG. 10. (a) Tunable cos-cos filter, (b) cos-sin filter, (c) Frequency responses $G_1(\omega)$ and $G_2(\omega)$ (without the scale factor j) are shown, (d) Frequency responses of the right and left filters, $H_R(\omega)$ and $H_L(\omega)$, obtained as sum and difference of $G_1(\omega)$ and $G_2(\omega)$ (Figure 11). The filters $H_R(\omega)$ and $H_L(\omega)$ are basically synthesized from a single prototype $H(\omega)$, and hence are perfectly matched and symmetric about ω_c. The frequency response of $H_C(\omega)$, not shown, is centered around ω_c. All filters are linear phase filters.

Further substituting for $H_1(\omega)$ and $H_2(\omega)$ in Eq. 23 from Eq. 20 and simplifying, we have

$$H_R(\omega) = H(\omega - \omega_c - \Delta) + H(\omega + \omega_c + \Delta))$$
$$H_L(\omega) = H(\omega - \omega_c + \Delta) + H(\omega + \omega_c - \Delta)). \quad (24)$$

Thus, the filters $H_R(\omega)$ and $H_L(\omega)$ (shown in Figure 10d) are the original prototype filter

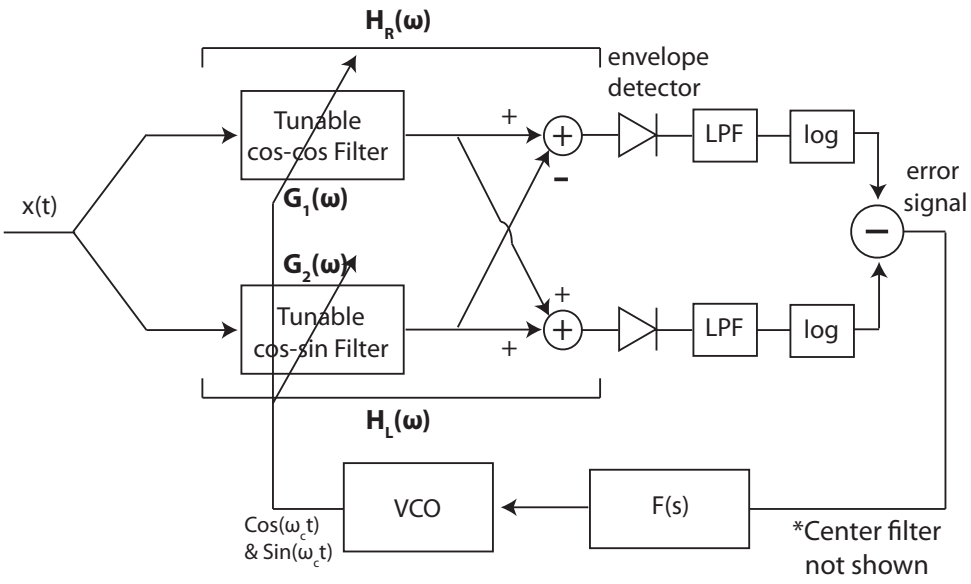

FIG. 11. Implementation of the frequency error detector and the frequency discriminator loop. The center filter $H_C(\omega)$ (not shown) is implemented using a cos-cos filter structure with $H(\omega)$ sandwiched between the multipliers as in Figure 5b.

$H(\omega)$ shifted to center frequencies $\omega_c + \Delta$ and $\omega_c - \Delta$, respectively. They have purely real valued frequency responses (except for the linear phase introduced by requiring a causal impulse response) and are the ones used in frequency error detection. In practice, the filter impulse responses in Eq. 19 are symmetrically truncated and Hann windowed about the time origin and made causal by shifting them to the right resulting in linear phase filters. The center filter $H_c(\omega)$ (also tunable) centered around ω_c, (shown in Figure 5b) is synthesized using the cos-cos structure, but with the prototype filter $H(\omega)$ sandwiched between the multipliers. Its output is not used in error signal calculation but is the channel output. If the input tone frequency ω_1 is less than the VCO frequency ω_c then the envelope at the output of $H_L(\omega)$ is larger than the envelope at the output of $H_R(\omega)$ and the error signal will drive the VCO to make ω_c equal to ω_1 and vice versa. The loop filter $F(s)$ determines the dynamics. The linear equivalent circuit described in section II.A is applicable to this implementation as well. The envelope detector shown in Figure 11 is a rectifier in cascade with a LPF. The logarithmic nonlinearity serves the same purpose as in DTF. This LPF

increases the time delay τ_g around the loop and has to be included while calculating the loop filter constants k_p and k_i.

III. SYNCHRONY CAPTURE FILTERBANK (SCFB)

The proposed synchrony capture filterbank (SCFB) shown in Figure 3a consists of a bank of fixed filters each cascaded with a frequency discriminator loop (FDL). The filterbank consists of K logarithmically spaced gammatone filters that have been widely used in auditory system modeling[32]. Using physiologically-appropriate filter parameters (approximately constant, low Q filters), gammatone filterbanks effectively replicate the broadly tuned mechanical filtering characteristics of the basilar membrane in the cochlea.

The gammatone filters used here were designed using the Auditory Toolbox developed by Malcolm Slaney[32], and further details of the cochlear model implementation are discussed in[33]. In our implementation K is 200. The constant-Q gammatone filters use a mix of "Glasberg and Moore" and "Lyon" parameters spanning center frequencies from 100-3940 Hz, with corresponding 3-db bandwidths ranging from 50 Hz to 905 Hz. Filter Q values (EarQ parameter) are all 4, and the order parameter is 1[33]. The minBW used in computing the equivalent rectangular bandwidth (ERB) is 50 Hz. The sampling frequency is 16000 Hz. An example of the frequency responses of one of the fixed filters and the associated three tunable filters of the SCFB are shown in Figure 12. Whereas the broadly tuned, fixed gammatone filters coarsely isolate the various frequency components in the incoming signal, the tunings of the more narrowly tuned bandpass triplet filters in the frequency discriminator loops (FDLs) converge on the precise frequencies of the individual frequency components.

A. Bandpass filter triplet parameters

As mentioned earlier each triplet of tunable filters consists of left, center, and right filters, $H_L(\omega)$, $H_C(\omega)$ and $H_R(\omega)$, whose center frequencies are spaced by a constant ratio. All of them are derived from a single prototype filter $H(\omega)$ defined in Eq. 18, whose frequency

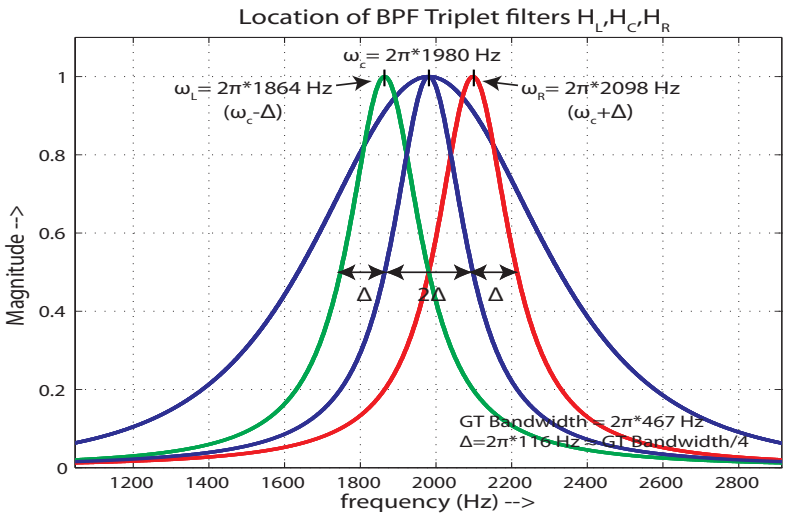

FIG. 12. A typical BPF Triplet centered at 1980 Hz. The broader frequency response corresponds to the gammatone filter centered around 1980Hz.

response is

$$H(\omega) = \frac{2\alpha}{\alpha^2 + \omega^2}. \tag{25}$$

The parameter α is chosen to be equal to the spacing between the filters, i.e., $\alpha = \Delta$. Δ has been chosen to be one-fourth of the bandwidth (actually halfwidth) of the gammatone filter. Hence $\alpha = \Delta = B_{GT}/4$ determines the prototype filter, where B_{GT} stands for gammatone filter bandwidth. For example, Figure 12 shows a gammatone filter centered around 1980 Hz with bandwidth of 466 Hz. Individual left, center and right triplet filters have center frequencies 1864, 1980, and 2098 Hz have bandwidths and center frequency spacings of 115 Hz. Bandwidths and spacings of fixed gammatone and adaptive triplet filters are proportional to center frequency.

B. Frequency discriminator loop filter design $F(s)$

The typical loop filter used in our implementation is of the form $F(s) = k_p + k_i/s$. The proportional gain k_p is intended to improve the rise time of the step response. The VCOs that steer the tuning of the triplet filters are initially set to match the center frequency

ω_c of their corresponding gammatone filter. Because the loop is initialized with the VCO frequency close to the input signal frequency, a consequence of the frequency selectivity of the associated gammatone filter, choosing $k_p = 0$ does not affect the loop's rise time performance significantly and also simplifies its implementation. On the other hand, k_i is needed to keep track of the frequency changes in the input and drive the steady state error to zero. The value of k_i depends on the frequency discriminator constant, k_s, and also on the parameter τ_g that represents the group delay of the prototype filter (i.e., its causal approximation) plus any delay introduced (in smoothing the envelope) in the envelope detector in Figure 11. For each channel, the following values were used for the loop filter parameters, and they seem to work well in most circumstances (set $\beta = 1$ in Eq. 7):

$$k_p = 0$$
$$k_i = \frac{1}{k_s}\left(21.90\frac{\gamma}{\tau_s^2}\right) = \frac{10.95\tau_g}{k_s\tau_s^2}.$$

τ_s, the settling time is chosen to be approximately $\frac{50}{f_c}$, where f_c is the center frequency of a gammatone filter. The FDL operation is not very sensitive to the choice of these parameters.

IV. SIMULATION RESULTS

The SCFB algorithm has been tested with appropriate parameter choices using several synthetic signals and speech signals drawn from the TIMIT database. Here simulation results are presented for one set of synthetic musical notes, an isolated utterance drawn from the ISOLET database, and a set of sentences of continuous speech from the TIMIT database with and without additive noise. For speech signals, the input signal is first subjected to spectral equalization by using a pre-emphasis filter and then processed through the filterbank and the self tuning FDL circuits. The frequencies of the VCOs in FDL modules indicate the frequency components that those modules are tracking and they are plotted as a function of time. The outputs of the BPF triplets are available for further processing, and these can be used to classify whether the signal in local frequency bands are tonal or noise-like. For example, if the envelope of the three filter outputs are larger than the background noise level

and if the center filter has a significantly larger output when compared with the associated left and the right filters, then this implies that the corresponding channel has a tonal signal. Conversely, if the three envelopes are approximately equal in size then this implies that the channel output is non-tonal or locally white.

A. Dyads of synthetic harmonic signals

The filterbank response to synthetic harmonic signals is considered first. The stimulus consists of two notes of two harmonic complexes (equal amplitude harmonics, 1 to 6). In musical terms, these are two notes separated by a a minor second (16:15) and a perfect fourth (4:3). They are the same signals that produced the auditory nerve interspike interval patterns depicted in Figure 2. The first note has two fundamentals (440 and 469 Hz) separated by 6.6%. The second has a frequency separation of 33.3% (with fundamental frequencies 440 and 587 Hz). Perceptually, for the minor second, human listeners hear only one pitch intermediate in frequency between the two notes, whereas for the perfect fourth, two note pitches can be heard.

Responses of the SCFB to these pairs of complex harmonic tones are shown in Figure 13. A "capturegram" plot of the resulting frequency tracks of the VCOs as a function of time shows the locking of groups of channels onto individual frequency components. The plots show only tracks of VCO frequencies of low frequency channels ($f_c < 1000$ Hz) to permit more direct comparison with the interspike interval histograms in Figure 2. Note that most of the VCO frequency tracks with CFs close to the dominant tone frequencies converge rapidly (within a few tens of milliseconds) to their steady state value.

The filterbank response for two closely spaced note dyads separated by 6.6% is shown in Figure 13a. This signal has 4 frequency components below 1000 Hz: 440, 469, 880, and 938 Hz. Here the filterbank does not resolve the pairs of nearby partials (440/469 and 880/938 Hz), but rather all the channels converge on the mean frequencies of the nearby partials (channels 53 to 88 fluctuate around 458 Hz, 89-112 fluctuate around 909 Hz). The

pattern of frequency capture is similar to that in the interspike interval data in Figure 2a. Figure 13b shows rectified outputs of each channel's center filter and Figure 13c shows the autocorrelation of the rectified outputs (from time t = 0.25 to 0.5 seconds). In this case we can see the fluctuations in envelope are related to the beat frequency (469-440=29 Hz) (as seen in Figure 2a).

The filterbank response to the well-separated note dyad is shown in Figure 13d. This signal has 3 frequency components below 1000 Hz: 440, 587, and 880 Hz. Clearly each VCO is captured by the dominant partial in that channel's neighborhood. Channels with center frequencies between 300 and 525 Hz lock to 440 Hz, those with center frequencies between 525 Hz and 725 Hz lock to 587 Hz, and the rest are captured by the 880 Hz partial. Transitions of VCO frequency change from one dominant tone to the other is abrupt. For example, for center frequencies near 500 Hz, the channels are either captured by 440 Hz tone or the 587 Hz tone. Very similar behavior is also observed in the interspike interval histograms in Figure 2b where interspike intervals in the corresponding CF channels switch abruptly from interval patterns associated with 440 Hz to those associated with 587 Hz. Figure 13e shows rectified outputs of each channel's center filter and Figure 13f shows the autocorrelation of the rectified outputs after the frequency estimates, which are almost constant (in other words the channel's VCO are locked, in this case from time = 0.25 to 0.5 seconds).

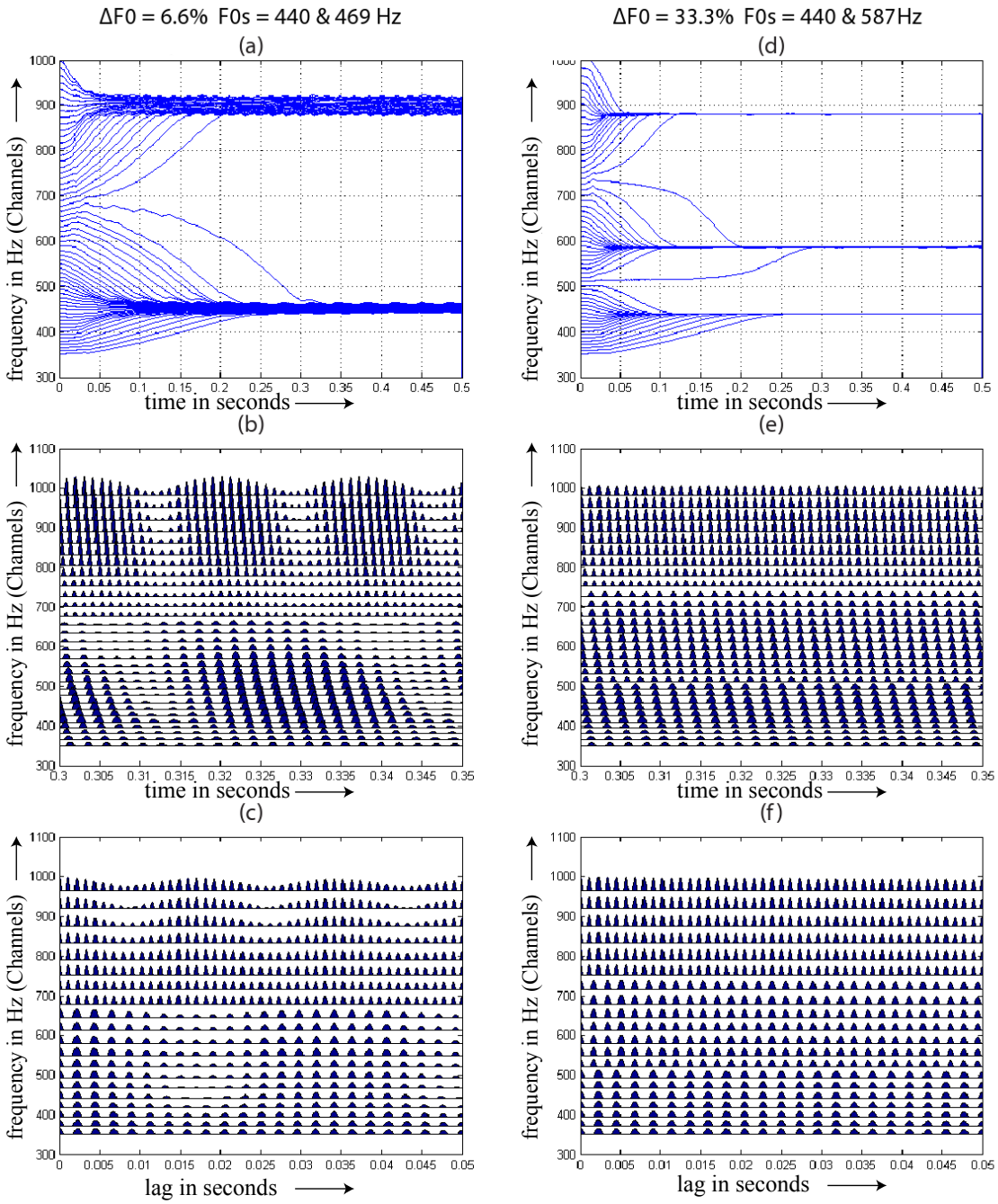

FIG. 13. Filterbank responses to pairs of harmonic tones. Left. Responses to a note dyad separated by a minor second (ΔF_0=6.6%, F_0s = 440 & 469 Hz). Right. Responses to a note dyad separated by a perfect fourth (ΔF_0=33.3%, F_0s = 440 & 587 Hz). Top plots (a),(d). Frequency tracks of the VCOs (capturegram). Middle plots (b), (e). Half-wave rectified output waveforms of channel center filters (analogous to a post-stimulus time neurogram). Bottom plots (c), (f). Channel autocorrelations (compare with autocorrelation neurograms of Figure 2).

B. Speech signals

For synthetic signals, such as the musical notes in the previous subsection, the instantaneous frequency estimates obtained from the VCOs of nearby channels are essentially the same after the initial settling time. However, for natural signals like speech the frequency estimates of the partials tend to have some variability (as can be seen below). Clearly, some sort of clustering method is needed to obtain the average frequency tracks associated with each frequency component in the signal. Other well known auditory-inspired models such as the ZCPA (Zero-Crossing Peak Amplitude)[34] or EIH (Ensemble Interval Histogram)[12] use the upward-going zero or level crossing events in a signal (emanating from a filter channel) to estimate the frequency. The reciprocal of the time interval between adjacent zero/level crossing events is used as the instantaneous frequency estimate. Such frequency estimates obtained over a time window are collected to assemble a frequency histogram. The frequency histograms across all filter channels are combined (in both ZCPA and EIH) to represent the output of the auditory model[34]. Further, in ZCPA the peak of the envelope that lies in between two consecutive zero-crossing events is used as a nonlinear weighting factor to a frequency bin to simulate the firing rate of the auditory nerve. In our case we follow a similar procedure except the frequency estimates are not derived from the zero-crossing events but from the VCOs frequencies. The envelopes are obtained from the rectified and smoothed outputs of the center filter of each channel.

The frequency values corresponding to the 200 channels are binned into 40 logarithmically spaced frequency bins that lie between 100 and 4000 Hz. However, before binning the frequency values, a non-linear weighting factor ($\log(1+a)$, where a is the amplitude/envelope corresponding to that frequency value) was applied as in ZCPA. Then the histogram peaks that have heights below a threshold (10% of the peak amplitude) are eliminated. This will eliminate the silence regions where the amplitudes are very low. Only when the log-envelope value is above the threshold, the actual frequency estimate of the frequencies in the bin are calculated using $\frac{\sum_n \log(1+a_n) f(n)}{\sum_n \log(1+a_n)}$, where a_n and f_n represent the amplitude/envelope

and frequency values that fall within a bin. The steps involved in the processing of speech signals are sketched in Figure 14a.

A histogram of the distribution of frequencies tracked by the VCOs is useful for assessing the degree to which channels have converged on particular frequencies. Here the number of channels converging on a particular frequency provides a robust, qualitative measure of its relative intensity. The running histogram of frequencies tracked (Figure 14a) provides a cleaner analysis of the time courses of dominant signal periodicities. Thresholding the running capture histogram keeps regions where multiples channels have converged on the same frequency and removes those where there is little agreement. Figures 14(b,c, and d), 15 and 16 demonstrate the character of this analysis.

C. Isolated spoken letters

The SCFB algorithm was applied to a vowel /i/ (as in "beet")(file name: fskes0-E1-t.adc, male speaker) drawn from the ISOLET database. Figure 14(b,c,d) shows the simulation results. Figure 14b shows the spectrogram of the vowel utterance and 14c shows the capturegram , i.e. the raw frequency tracks of the 200 VCOs.

It can be seen that the FDLs track closely the frequencies of the individual partials up to at least 1000 Hz. Depending on the relative intensity of each partial, typically five to ten channels tend to converge on to the stronger partials' frequency tracks. The first formant F_1 is located at around 300 Hz between the second and third harmonics. At higher frequencies (> 2000 Hz), where the filters (the gammatone and BPFs tend to be wider) several channels tend to converge on the three higher formant frequencies which are located approximately at frequencies 2400, 2800 and 3800 Hz. Between the first and the second formant frequencies where the signal energy is relatively low there are no dominant tones and hence the VCO tracks tend to wander. Figure 14d shows the cleaned up tracks after the histogramming procedure outlined in Figure 14a is applied. This procedure tends to suppress meandering tracks and signal components with small envelope values.

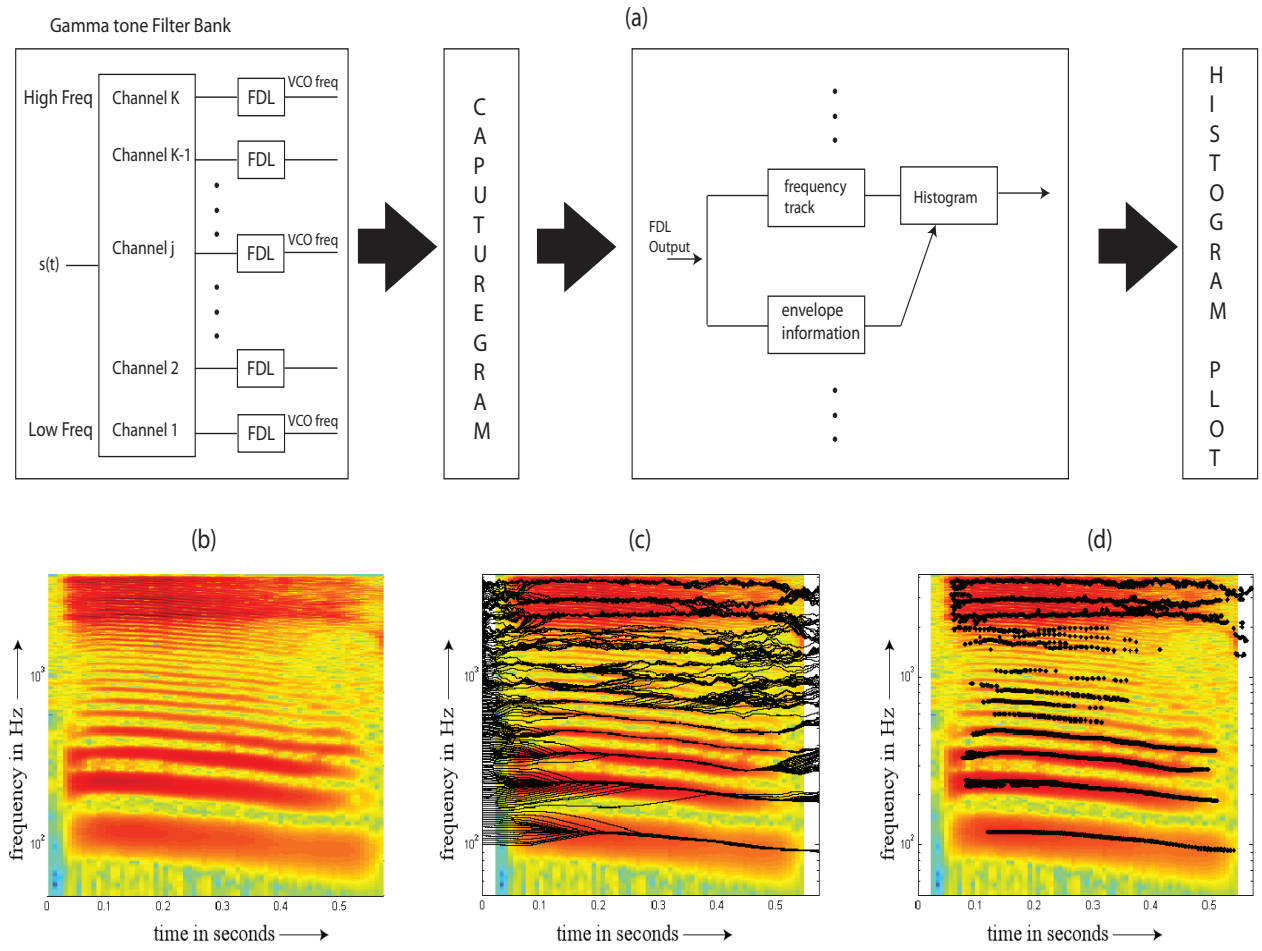

FIG. 14. (a) Steps involved in the SCFB algorithm. The input speech signal s(t) (after preemphasis) is processed by the 200 gammatone filters and the associated FDLs and the frequency tracks are plotted as capturegrams. The VCO frequency values and the associated envelopes are used to generate the frequency histograms from which dominant frequency tracks are derived. Results for ISOLET vowel /i/. (b) Spectrogram (c) Capturegram (d) Thresholded histogram plot.

D. Continuous speech

The SCFB algorithm was also applied to several continuous speech samples drawn from the TIMIT database. The speech signals were first pre-emphasized with a $H(z) = 1-0.95z^{-1}$ filter to equalize the spectrum to prevent strong low frequency components from swamping the weaker high frequency components. The sampling frequency is 16kHz. Capturegrams for

two speech sentences, "Where were you while we were away?" (TIMIT sx9) and "The oasis was a mirage" (TIMIT sx280) spoken by male and female speakers are shown in Figures 15 and 16, respectively.

Figures 15a and 15d show the spectrograms of the TIMIT sx9 utterances by male and female speakers. In Figure 15b and 15e the corresponding capturegram tracks for the 200 VCOs are superimposed on the spectrogram for the male and female utterances. Typically, for a strong low-frequency harmonic component, a handful of channels are captured by one harmonic. Note that at low frequencies and harmonic numbers (f $<$ 800 Hz, n $<$ 8) almost all the individual harmonics tend to be closely tracked by the FDLs. These frequency tracks together can provide a robust representation of the fundamental frequency (voice pitch). For higher frequencies and harmonic numbers, only dominant harmonics in formant regions are tracked. This behavior is due to the constant Qs of the filters, such that FDL triplet filters with higher center frequencies have correspondingly larger bandwidths, and therefore cannot resolve individual harmonics. Instead these filters lock onto the nearest dominant harmonic component somewhere near the middle of a formant.

Similarly, Figures 16b and 16e show the capturegrams for the sentence TIMIT sx280 spoken by a male and a female, respectively. In both cases, the frequency transitions, especially at the higher frequency regions are precisely and robustly tracked. At lower frequencies, as one harmonic becomes weaker with respect to a nearby harmonic, the frequency tracks of channels in that neighborhood jump from the weaker harmonic to the stronger one due to the tendency of the FDL to track the stronger component (as in the time-frequency region t = 1.0 -1.45 s, frequency $<$ 1000 Hz) in Figure 16e. Again the last rows of both figures show the tracks after the histogramming procedure is used to clean up the raw tracks data.

Previous analysis of cat auditory nerve responses had suggested that the synchrony capture effect is resistant to noise[35]. So, we tested the SCFB algorithm with noisy speech signals to determine its robustness to noise. Signal power P_s is calculated as the sum of squares of all the speech signal samples divided by the time duration of the speech signal. The variance σ^2 is obtained from the definition of signal to noise ratio (SNR) given below.

$$SNR = 10 \log_{10} \left(\frac{P_s}{\sigma^2} \right) dB. \tag{26}$$

The Gaussian distributed noise samples are generated with a variance σ^2 obtained from the above formula for an SNR of 10 dB. The generated noise samples are added to the speech signals, and are processed by the SCFB algorithm. Figure 17 shows the simulation results. Left column corresponds to "The oasis was a mirage" (sx280) for a female speaker, and the right column is for "Where were you while we were away?" (sx9) by a male speaker. The spectrograms (a) and (d) are relatively darker than the spectrograms in Figures 15 and 16, because of the added 10dB noise. Even in these noise corrupted cases, the formant and harmonics' tracks (especially the formant transitions) are clearly visible. Capturegrams show that multiple channels still merge to the same frequencies and the histogram tracks are also relatively clean. Thus the behavior of the SCFB in noise seems to parallel that seen in the cat auditory nerve.

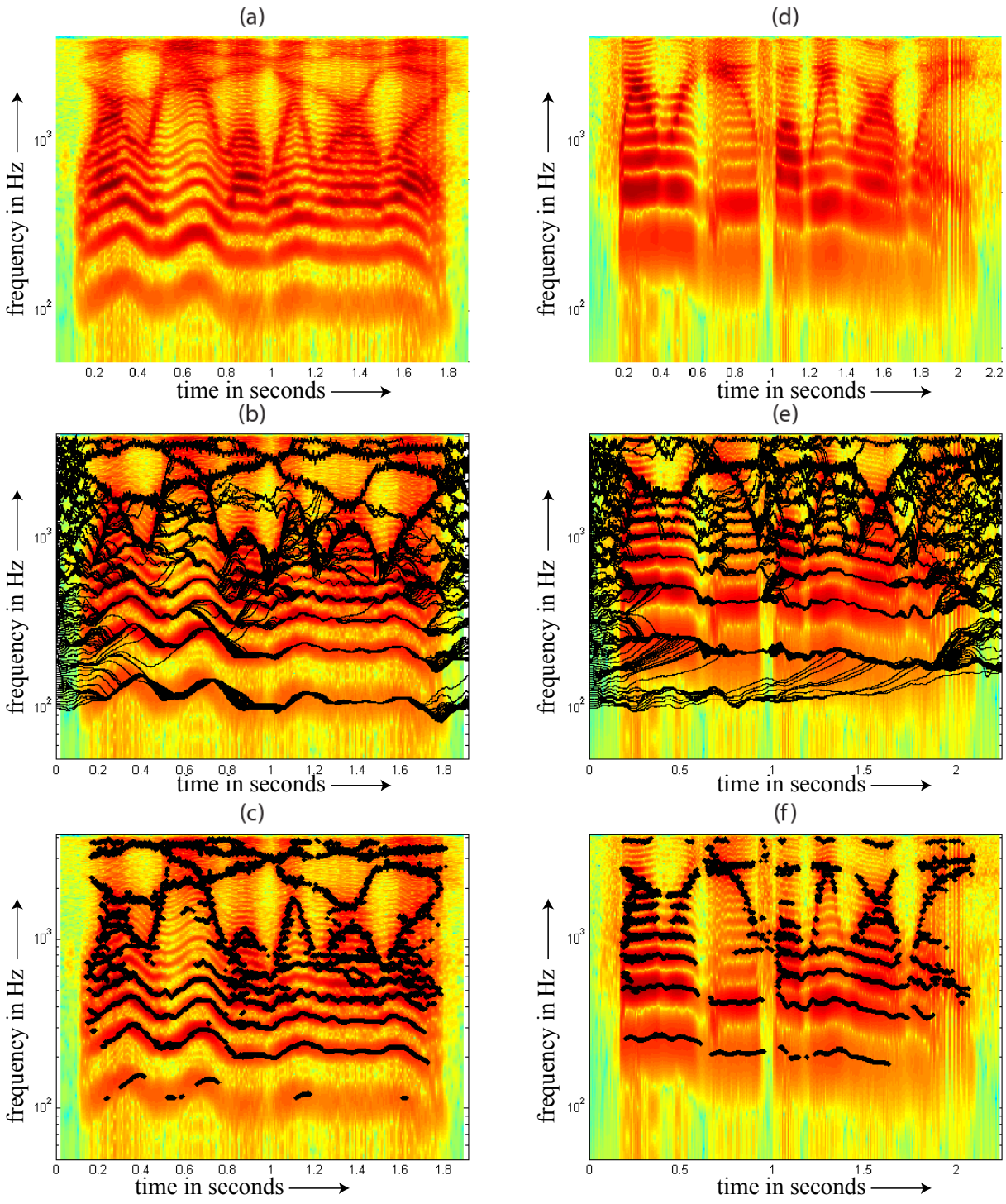

FIG. 15. Results for TIMIT utterance, "Where were you while we were away?" (sx9) for male (left column) and female (right column) speakers. Top plots (a)(d). Spectrograms. Middle plots (b)(e). Capturegrams. Bottom plots (c)(f). Thresholded histogram plots. At low frequencies, all individual harmonics are tracked, whereas above 1000 Hz, only prominent formant harmonics are tracked.

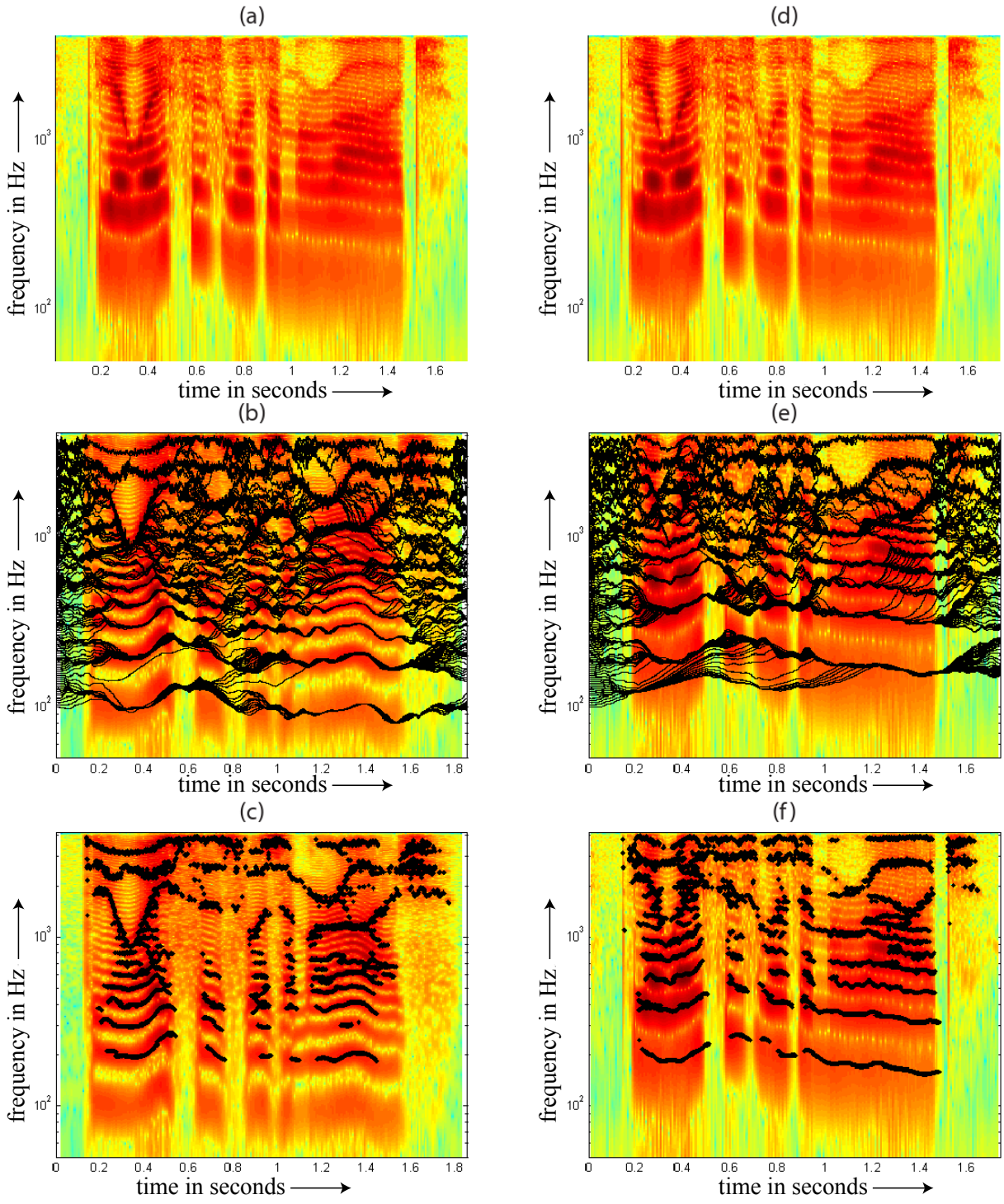

FIG. 16. Results for TIMIT utterance "The oasis was a mirage" (sx280) for male (left column) and female (right column) speakers. Plots as in the previous figure. High frequency frication above 4000 Hz in "oasis" not shown.

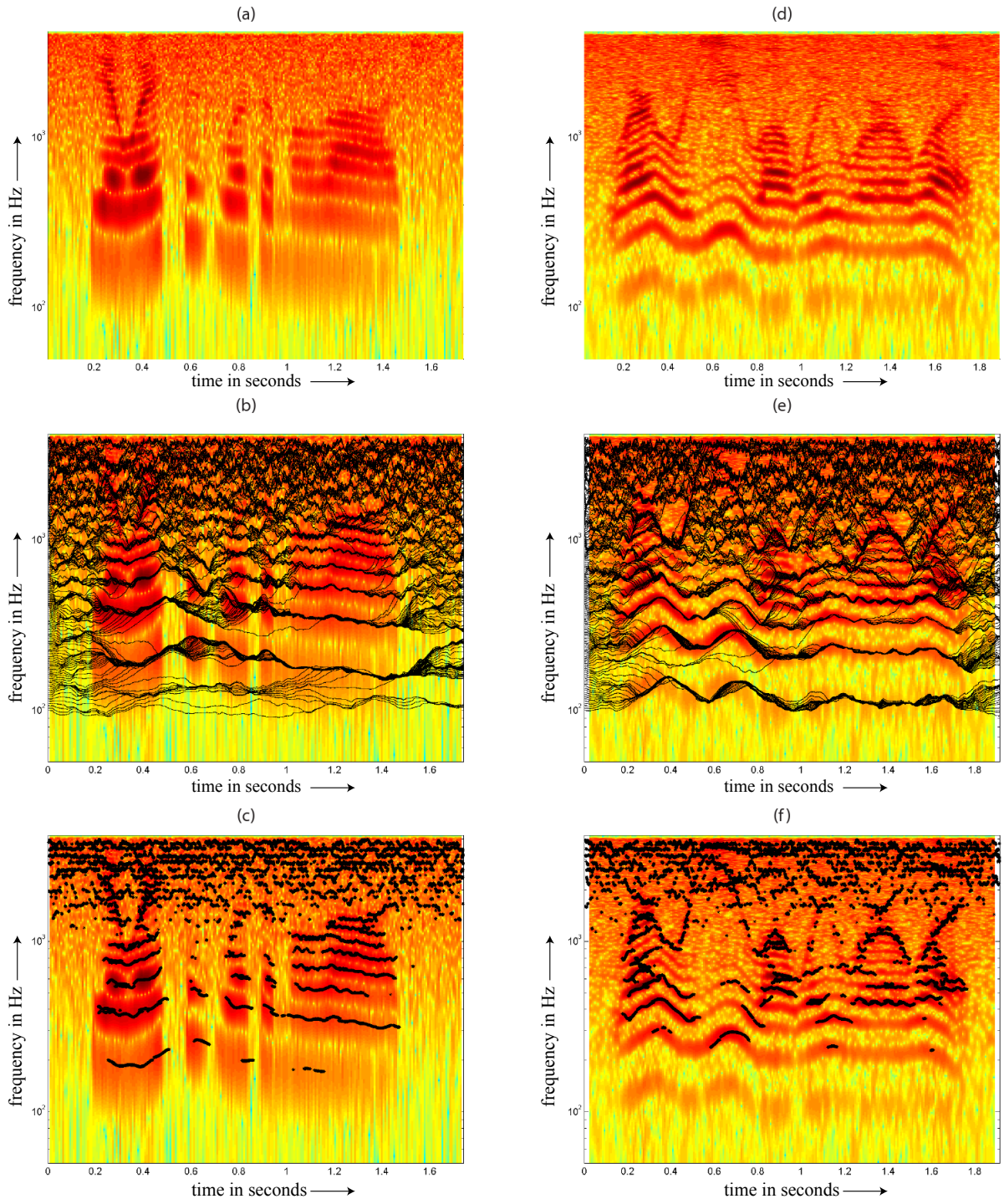

FIG. 17. Results for two TIMIT utterances in 10dB noise. "The oasis was a mirage" (sx280) for a female speaker (left column) and "Where were you while we were away?" (sx9) for a male speaker (right column). Plots as in the previous figure.

V. DISCUSSION

Our interest in synchrony-capture based filterbanks has been motivated by considerations of the functional anatomy and response characteristics of the cochlea, adaptive filtering signal processing strategies in radar and other artificial systems, and the possible role of synchrony capture in auditory nerve representation of complex sounds. The primary goal in this first stage of investigation has been to integrate these aspects into a workable algorithm for tracking the major frequency components present in an acoustic signal.

A. Relationship to previous signal processing strategies

As is often the case, the signal processing constituents of the SCFB algorithm proposed here have a long history. Frequency discriminator loops (FDLs) have been used in digital and analog communication systems for signal tracking for many decades[27]. The frequency error detector (FED) circuit (Figure 4) is a key component of the FDL that senses the difference between the frequency of the input signal and that of a local VCO in order to produce a proportional error voltage that can be used for steering purposes.

Basically there are two or three common types of frequency error detector circuits that are used in practice. The quadricorrelator[28,29], briefly outlined in Appendix A, is often used in communication systems. The other type, which has been used here in the SCFB design, uses stagger-tuned filters and compares envelopes of filter outputs to derive running error voltages. Ferguson and Mantey[21] originally proposed the use of such adaptable stagger-tuned bandpass filters for frequency error detection. Alternately, frequency error detectors can also be implemented directly by using phase derivatives of a complex signal (see for example[36,37]). Wang[38] has designed a harmonic locked loop to track the fundamental frequency of a periodic signal using this idea. However, these approaches require a complex (Hilbert-transformed) signal for processing.

In their adaptive, stagger-tuned design, Ferguson and Mantey used the error voltage (envelope difference) to retune the bandpass filters directly by moving their pole locations.

Such a design does not use VCOs to tune the filters. Based on this idea one could imagine cochlear filters where the frequency response of a filter is adjusted by changing a mechanical parameter such as stiffness depending on the envelope voltage difference between the left and the right filters. Costas[22] used a similar FED, but used the error voltage to change the frequency of a VCO that indirectly moved the left and the right bandpass filters in tandem. The proposed approach is closer to Costas' method and its variants[22,36,38]. The main difference is that a compressive (logarithmic) nonlinearity is used on the envelope of a signal to suppress nearby weaker signal components. Such compressive nonlinearities have the property of favoring a stronger component in the presence of other weaker ones. This is the primary reason that synchrony capture occurs.

The SCFB design is also related to adaptive formant tracking methods proposed earlier by Rao and Kumaresan[39,40], and subsequently improved by Mustafa and Bruce[41]. However, in Rao-Kumaresan approach the adaptive formant filters were controlled by measuring the instantaneous frequency of a complex-valued signal. Further, as mentioned earlier, EIH and ZCPA algorithms also estimate the frequency of tonal signals based on the zero or level crossing intervals. However, these may be regarded as open loop methods for estimating instantaneous frequencies, unlike the closed loop methods like FDL.

B. Similarities to response characteristics of the cochlea and auditory nerve

Although the SCFB is not a biophysical model, its signal processing behavior bears many qualitative similarities to response patterns in the mammalian cochlea. First, the mammalian cochlea produces acoustic emissions, called spontaneous otoacoustic emissions (SPOAEs)[42]). The narrow spectral widths of these emissions suggest that they are generated by spontaneous oscillations in the cochlea, possibly in outer hair cells. This kind of behavior is also characteristic of voltage controlled oscillators that implement the FDL in the present architecture.

Second, it is also well known[42] (p.117) that the cochlea also produces acoustic emissions

at additional frequencies when two tones of frequency f_1 and f_2 ($f_2 > f_1$) are presented. Listeners can often hear discordant faint tones not present in the original stimulus. The strongest of these cochlear distortion products, the cubic distortion product generated at $2f_1 - f_2$ Hz, is thought to be a direct byproduct of cochlear mechanics, in the form of a compressive nonlinearity in OHC response. The ensuing signal distortions are analogous to intermodulation products in communication systems. The FDL architecture produces similar combination tones as a byproduct of its operation. Consider the operation of the FDL as described in section II.B when two simultaneous tones with frequencies f_1 and f_2 and corresponding amplitudes A_1 and A_2 are applied as input. The spectrum of the VCO output for this stimulus is shown in Figure 18 for a channel with center frequency 1890 Hz. $f_1 = 1950$ Hz and $f_2 = 2050$ Hz, $A_1 = 1$ and $A_2 = 0.5$. Note that the VCO locks on to the stronger tone at f_1 Hz and that the left and the right filters of that channel adjust themselves such that their average envelopes are equal. Then the resulting error signal $e(t)$ is proportional to $C\cos(\Delta\omega t)$ where $\Delta\omega = 2\pi \times (f_2 - f_1)$ and C is a constant related to the ratio of amplitudes A_2/A_1 (see Eq. 14). This error signal then frequency modulates the VCO's carrier at the dominant tone frequency f_1. The resulting frequency modulated VCO output has sideband components at $f_1 \pm n(f_2 - f_1)$[18] p.180-87. The output spectrum in Figure 18 shows some of the sidebands (for $n = 1$ and 2). Thus qualitative parallels exist between combination tones produced by live cochleae and the VCO-driven frequency capture circuits of the filterbank.

Two-tone suppression is a third nonlinear phenomenon. Like the cochlea, the proposed filterbank produces both rate- and synchrony-suppression. Two-tone rate suppression is generally regarded as a nonlinear property of the cochlea in which the average neural firing rate in the region most sensitive to a probe tone is reduced by the addition of a suppressor tone at a different nearby frequency. For the filterbank, when dominant frequency components steer the tunings of local VCOs away from other frequencies, responses to less intense secondary tones at those frequencies are attenuated relative to those produced when the dominant tone is absent.

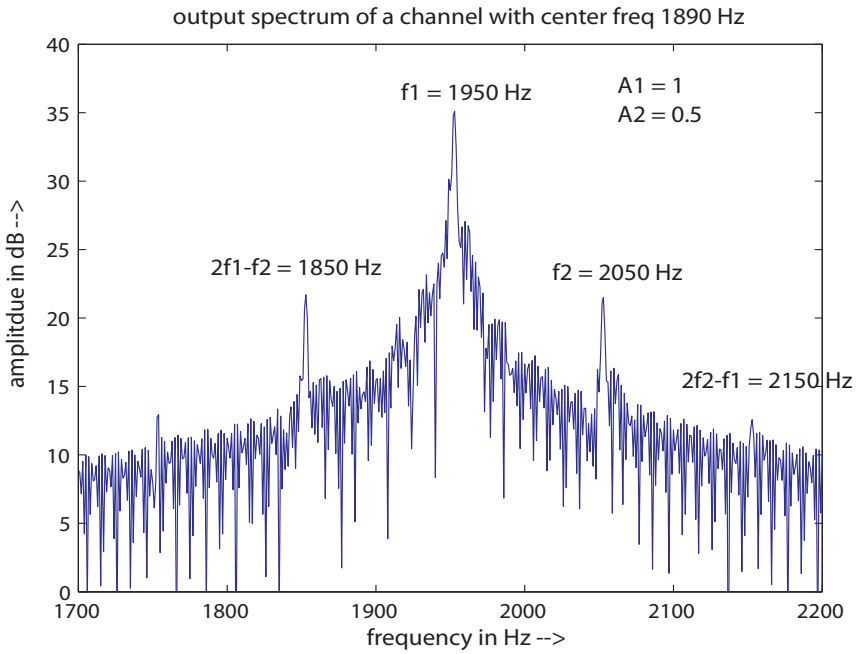

FIG. 18. Distortion products. Spectrum of VCO output signal of a channel with center frequency of 1890 Hz in response to two pure tones at frequencies $f_1 = 1950$ Hz and $f_2 = 2050$ Hz with amplitudes $A_1 = 1$ and $A_2 = 0.5$ respectively. Note occurrences of distortion products at frequencies $f_1 \pm n(f_2 - f_1)$. These are generated in frequency discriminator loops when VCOs lock on to dominant tones at f_1 but are also frequency modulated by an error signals consisting of a weak tones at $\Delta f = f_2 - f_1$.

There is also the related phenomenon of synchrony suppression. The effects of two tonal inputs on temporal patterns of neural firing have been extensively studied. Auditory nerve fibers phase-lock in response to low frequency tones (< 5000 Hz), i.e. spikes are mainly produced at particular phase angles of the waveform[11]. The degree of synchronization of spikes to a given frequency can be quantified by computing the vector strength ("synchronization index") of the spike distribution as a function of waveform phase. When the stimulus consists of two tones, Hind et al.[43] found that auditory nerve spikes may be phase locked to one tone, or to the other, or to both tones simultaneously. Which of these occurs is determined by the relative intensities of the two tones and their frequencies and spacings. Moore[11] summarizes these results as follows, "When phase locking occurs to only one tone of a pair,

each of which is effective when acting alone, the temporal structure of the response may be indistinguishable from that which occurs when the tone is presented alone. Further, the discharge rate may be similar to the value produced by that tone alone. Thus the dominant tone appears to "capture" the response of the neuron. This (synchrony) capture effect underlies the masking of one sound by another". The tone that is suppressed ceases to contribute to the pattern of phase-locking, and the neuron responds as if only the suppressing tone were present. The effect is that the synchronization index of a fiber to a given tone is reduced by the application of a second tone[44]. Similarly, in the filterbank, capture of a given channel VCO by a locally dominant component produces an output waveform having the frequency of the dominant tone, causing the vector strength of the dominant component to increase at the expense of those of weaker secondary ones.

VI. CONCLUSIONS

A striking feature of the phase-locked responses to complex sounds is the phenomenon of "synchrony capture"[3,5], wherein an intense stimulus frequency component dominates the temporal firing patterns of auditory nerve fibers innervating the corresponding cochlear frequency region. The capture effect refers to the almost exclusive nature of the phase-locking to the dominant component, such that the output of whole subpopulations of auditory nerve fibers in a cochlear region respond in the same way. Synchrony capture may be critical for separation of concurrent harmonic sounds.

An adaptive filterbank structure is proposed that emulates synchrony capture in the auditory nerve. This filterbank has two parts: a fixed array of traditional, passive linear (gammatone or equivalent) filters that are cascaded with a bank of adaptively tunable bandpass filter triplets. Envelope differences in the outputs of the filters that form the triplets are used in frequency discriminator loop (FDL) to steer their center frequencies with the help of a voltage controlled oscillator (VCO).

The resulting filterbank exhibits many desirable properties for processing speech and

other natural sounds. First, the number of channels converging on a particular frequency yields a robust means of encoding the intensity of the driving frequency component. The VCOs track resolved harmonics, which are known to be essential in determining the pitch and for the separation of concurrent periodic sounds. For voiced speech, the VCOs track the strongest harmonic in each formant region, yielding precise features for formant tracking.

VII. ACKNOWLEDGMENTS

This work was supported by the Airforce Office of Scientific Research under the grant # AFSOR FA9550-09-1-0119. We thank Prof. R. Vaccaro for pointing out that the Laplace transform of a time delay operator $\delta(t - \tau_g)$ $(= e^{-s\tau_g})$ can be approximated (Padé approximation) by a ratio of s-polynomials. The authors thank the three reviewers for many suggestions that helped improve the manuscript.

VIII. APPENDIX A: ALTERNATE FREQUENCY ERROR DETECTORS

The frequency error detector (FED) is a key component of the FDL (see Figure 4). In the tone followers described in section II we used the difference in (squared) envelopes (or log-envelopes) of the outputs of $H_R(\omega)$ and $H_L(\omega)$ as the error signal $e(t)$. $e(t)$ is proportional to the difference between the VCO frequency ω_c and the input (or dominant) tone frequency ω_1. In section II the specific type of FED (that is, one that uses squared envelope differences) was chosen because of its apparent functional similarity to the functioning of cochlear hair cells. (The inner/outer hair cells act as halfwave rectifiers followed by low-pass filters). Disregarding such constraints, if computer implementation of a FDL is the primary goal, then many other FEDs are available. Of course, the frequency error signal could be positive or negative depending on whether ω_c is greater or smaller than ω_1. Therefore, any method that is used to measure the frequency of a single tone can serve as a FED as long as it is also capable of detecting the sign of the frequency error. One such FED is called a Quadricorrelator[28]. The quadricorrelator (refer to Figure 3 in[28]) is input with a tone $A_1 \cos(\omega_1 t + \theta_1)$ and the VCO outputs $\cos(\omega_c t)$ and $\sin(\omega_c t)$. The low pass filters (LPF) (in Figure 3 in[28]) retain only the difference frequency outputs $\alpha_1 \cos(\Delta\omega t + \theta_1)$ and $\alpha_2 \sin(\Delta\omega t + \theta_1)$. The two differentiator outputs after cross multiplying (in Figure 3 in[28]) are added together to produce the error signal which retains the sign of the frequency error. Since in our simulations, in-phase and quadrature-phase signals (I and Q) are available, complex valued processing can also be used to estimate frequency error[37,38,45].

IX. APPENDIX B: EXPRESSIONS FOR THE FREQUENCY DISCRIMINATOR CONSTANT k_s

k_s, defined in section II.A, is the slope of the frequency discriminator function $S(\omega)$ at ω_c. $S(\omega)$ for the Simple Tone Follower (STF) is defined as

$$S(\omega) = \frac{|H_R(\omega)|^2 - |H_L(\omega)|^2}{|H_R(\omega)|^2 + |H_L(\omega)|^2} \tag{27}$$

where $|H_R(\omega)|^2 = |H(\omega - (\omega_c + \Delta))|^2$ and $|H_L(\omega)|^2 = |H(\omega - (\omega_c - \Delta))|^2$. Using $H(s) = \frac{1}{s+\alpha}$, $H(\omega) = \frac{1}{j\omega + \alpha}$, $|H_R(\omega)|^2$ and $|H_L(\omega)|^2$ are

$$|H_R(\omega)|^2 = \frac{1}{(\omega - (\omega_c + \Delta))^2 + \alpha^2} \tag{28}$$

$$|H_L(\omega)|^2 = \frac{1}{(\omega - (\omega_c - \Delta))^2 + \alpha^2} \tag{29}$$

Substituting Eqs. 28 and 29 in Eq. 27, we get

$$S(\omega) = \frac{2\Delta(\omega - \omega_c)}{\omega^2 + \omega_c^2 + \Delta^2 - 2\omega\omega_c + \alpha^2}. \tag{30}$$

k_s is obtained by taking the derivative of $S(\omega)$ with respect to ω and evaluating at $\omega = \omega_c$.

$$k_s = \left[\frac{dS(\omega)}{d\omega}\right]_{\omega=\omega_c} = \frac{2\Delta}{\Delta^2 + \alpha^2}. \tag{31}$$

Similarly, for the Dominant Tone Follower (DTF), k_s is obtained by taking the derivative of $S(\omega) = \log\frac{|H_R(\omega)|^2}{|H_L(\omega)|^2}$ and evaluating at $\omega = \omega_c$. It is easy to show that

$$k_s = \frac{4\Delta}{\Delta^2 + \alpha^2}. \tag{32}$$

References

[1] R. M. Stern and N. Morgan, "Hearing is believing," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 34–43, 2012.

[2] M. Sachs and E. Young, "Encoding of steady-state vowels in the auditory nerve: representation in terms of discharge rate," *J Acoust Soc Am*, vol. 66, pp. 470–479, 1979.

[3] B. Delgutte and N.Y.S. Kiang, "Speech coding in the auditory nerve: I. Vowel-like sounds," *J. Acoust. Soc. Am.*, vol. 75, pp. 866–878, 1984.

[4] H.E. Secker-Walker and C.L. Searle, "Time-domain analysis of auditory-nerve-fiber firing rates," *J. Acoust. Soc. Am.*, vol. 88, no. 3, pp. 1427–1436, 1990.

[5] M.B. Sachs, I.C. Bruce, R.L. Miller, and E.D. Young, "Biological basis of hearing-aid design," *Annals of Biomedical Engineering*, vol. 30, pp. 157–168, 2002.

[6] M. J. Tramo, P. A. Cariani, B. Delgutte, and L. D. Braida, "Neurobiological foundations for the theory of harmony in western tonal music," *Ann N Y Acad Sci*, vol. 930, pp. 92–116, 2001.

[7] P. Cariani and B. Delgutte, "Neural correlates of the pitch of complex tones.I. Pitch and pitch salience. II. Pitch shift, pitch ambiguity, phase-invariance, pitch circularity, and the dominance region for pitch," *J. Neurophysiol.*, vol. 76(3), pp. 1698–1734, 1996.

[8] R. Kumaresan, V. K. Peddinti, and P. Cariani, "Synchrony capture filterbank (SCFB): An auditory periphery-inspired method for tracking sinusoids," in *Proceedings of the ICASSP*, Kyoto, Japan, 2012, pp. 153–156.

[9] P. Cariani, "Temporal coding of periodicity pitch in the auditory system: an overview," *Neural Plasticity*, vol. 6, no. 4, pp. 147–172, 1999.

[10] R. Meddis and L. O'Mard, "A unitary model of pitch perception," *J Acoust Soc Am*, vol. 102, no. 3, pp. 1811–20, 1997.

[11] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, pp. 40, 90–103, 247, Academic Press, San Diego, fourth edition, 1997.

[12] O. Ghitza, "Auditory models and human performance in tasks related to speech coding

and speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 2, pp. 115–132, 1994.

[13] L. Cedolin and B. Delgutte, "Spatiotemporal representation of the pitch of harmonic complex tones in the auditory nerve," *J. Neurosci.*, vol. 30(6), pp. 12712–12724, 2010.

[14] C. J. Darwin, "Auditory grouping," *Trends in Cognitive Sciences*, vol. 1, no. 9, pp. 327–333, 1997.

[15] C.J. Plack and A. J. Oxenham, "Psychophysics of Pitch," in *Pitch: Neural Coding and Perception*, C.J. Plack, A.J. Oxenham, R. R. Fay and A.N. Popper, Ed., Springer Handbook of Auditory Research, chapter 2, pp. 7–55. Springer-Verlag, New York, 2005.

[16] J. O. Pickles, "Psychophysical frequency resolution in the cat as determined by simultaneous masking and its relation to auditory-nerve resolution," *J. Acoust. Soc. Am.*, vol. 66, pp. 1725–1732, Dec. 1979.

[17] E. J. Baghdady, "Theory of stronger-signal capture in FM reception," in *Proceedings of Institute of Radio Engineers*, Aalborg, Denmark, April 1958, pp. 728–738.

[18] S. Haykin, *Communication Systems*, John Wiley & Sons, NY, second edition, 1983.

[19] Luis Robles and Mario A. Ruggero, "Mechanics of the mammalian cochlea," *Journals of the American Physiological Society*, vol. 81, no. 3, pp. 1305–1352, July 2001.

[20] W.E. Brownell, "The Piezoelectric Outer Hair Cell," in *Vertebrate Hair Cells*, R.A. Eatock, R.R. Fay and A.N. Popper, Ed., Springer Handbook of Auditory Research, chapter 7, pp. 313–347. Springer-Verlag, New York, 2010.

[21] M.J.Ferguson and P.E.Mantey, "Automatic frequency control via digital filtering," *IEEE Trans. on Audio and Electroacoustics*, vol. AU-16, no. 3, pp. 392–397, Sep 1968.

[22] J. P. Costas, "Residual signal analysis -a search and destroy approach to spectral estimation," in *Proceedings of the First ASSP Workshop on Spectral Estimation*, Hamilton, Canada, Aug. 1981, pp. 6.5.1–6.5.8.

[23] P. Dallos, "Overview: Cochlear Neurobiology," in *The Cochlea*, P. Dallos, A. N. Popper, and R. R. Fay, Ed., Springer Handbook of Auditory Research, chapter 1, pp. 1–43. Springer-Verlag, New York, 1996.

[24] Aage R. Møller, *Hearing: Its Physiology and Pathophysiology*, Academic Press, first

edition, 2000.

[25] Fabio A. Thiers, Joseph B. Nadol, and M. Charles Liberman, "Reciprocal synapses between outer hair cells and their afferent terminals: Evidence for a local neural network in the mammalian cochlea," *J. Assoc. Res. Otolaryngol.*, vol. 9, pp. 477–489, 2008.

[26] H. Spoendlin, "Degeneration behaviour of the cochlear nerve," *Archiv fr klinische und experimentelle Ohren-, Nasen- und Kehlkopfheilkunde*, vol. 200, pp. 275–291, 1971.

[27] F. D. Natali, "AFC tracking algorithms," *IEEE Trans. Commun.*, vol. Com-32, pp. 935–947, Aug. 1984.

[28] F. M. Gardner, "Properties of frequency difference detectors," *IEEE Transactions on Communications*, vol. Com-33, no. 2, pp. 131–138, 1985.

[29] D. G. Messerschmitt, "Frequency detectors for PLL acquisition in timing and carrier recovery," *IEEE Trans. Comm.*, vol. COM-27, no. 9, pp. 1288–1295, Sep 1979.

[30] R. J. Vaccaro, *Digital Control: A State-Space Approach*, chapter 6, pp. 233, 254, McGraw-Hill, NY, 1st edition, 1995.

[31] G. H. Golub and C. F. Van Loan, *Matrix Computations*, chapter 11, pp. 572–574, The Johns Hopkins University Press, Baltimore and London, third edition, 1996.

[32] J. Holdsworth, I. Nimmo-Smith, R. Patterson, and P. Rice , "Implementing a gamma-tone filter bank," in *Annex C of the SVOS Final Report (Part A: The Auditory Filter Bank) MRC (Medical Research Council), APU (Applied Psychology Unit) Report 2341.* University of Cambridge, Cambridge, United Kingdom, Feb 1988.

[33] M. Slaney, "An efficient implementation of the patterson-holdsworth auditory filter bank," in *Apple Technical Report # 35, Perception Group - Advanced Technology Group*, Apple Computer Library, Cupertino, CA 95014, 1993.

[34] D. Kim, S. Lee, and R. Kil, "Auditory processing of speech signals for robust speech recognition in real-world noisy environments," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 1, pp. 55–69, Jan. 1999.

[35] M. B. Sachs, H. F. Voigt and E. D. Young, "Auditory nerve representation of vowels in background noise," *J. Neurophysiol.*, vol. 50, no. 1, pp. 27–45, Jul. 1983.

[36] R. Kumaresan, C. S. Ramalingam, and A. Rao, "RISC: an improved costas estimator-predictor filter-bank for decomposing multi-component signals," in *Proceedings of the Seventh Statistical Signal and Array Processing Workshop*, Québec City, Canada, June 1994, pp. 207–210.

[37] S. M. Kay, "A fast and accurate single frequency estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, pp. 1987–1990, Dec. 1989.

[38] A.L.Wang, *Instantaneous and frequency warped signal processing techniques and auditory source separation*, Ph.D. thesis, Stanford University, Stanford,CA, August 1994.

[39] R. Kumaresan and A. Rao, "Model-based approach to envelope and positive-instantaneous frequency of signals and application to speech," *Journal of the Acoustical Society of America*, vol. 105 (3), pp. 1912–1924, March 1999.

[40] A. Rao and R. Kumaresan, "On decomposing speech into modulated components," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 3, pp. 240–254, May 2000.

[41] K. Mustafa and I. C. Bruce, "Robust formant tracking for continuous speech with speaker variability," *IEEE Trans. on Speech Audio Processing*, vol. 14, no. 2, pp. 435–444, 2006.

[42] P. A. Fuchs, "Otoacoustic emissions and evoked potentials," in *The oxford handbook of auditory science: The ear*, David T Kemp, Ed., chapter 4, pp. 93–137. Oxford University Press, Great Clarendon Street, Oxford OX26DP, 1st edition, 2010.

[43] J.E.Hind, D.J.Anderson, J.F.Brugge, and J.E.Rose, "Coding of information pertaining to paired low-frequency tones in single auditory nerve fibers of the squirrel monkey," *Journal of Neurophysiology*, vol. 30, pp. 794–816, July 1967.

[44] E. Javel, C. D. Geisler, and A. Ravindran, "Two-tone suppression in auditory nerve of the cat: Rate-intensity and temporal analyses," *J. Acoust. Soc. Am.*, vol. 63, pp. 1093–1104, 1978.

[45] R. Kumaresan and C. S. Ramalingam, "On separating voiced-speech into its components," in *Proceedings of the Twenty-Seventh Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, Nov. 1993, pp. 1041–1046.

List of Figures